Mixture Multigroup Bayesian SEM with approximate measurement invariance for comparing structural relations across many groups

Hongwei Zhao[1], Jeroen K. Vermunt[2], and Kim De Roover[1,2]

[1]KU Leuven

[2]Tilburg University

Author Note

Hongwei Zhao https://orcid.org/0009-0008-9372-4986

Jeroen K. Vermunt https://orcid.org/0000-0001-9053-9330

Kim De Roover https://orcid.org/0000-0002-0299-0648

Funded by the European Union (ERC, PROCESSHETEROGENEITY, 101040754, awarded to Kim De Roover). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The code behind the project has been made publicly available on https://osf.io/rtp78/.

Correspondence concerning this article should be addressed to Hongwei Zhao, Quantitative Psychology and Individual Differences, KU Leuven, Tiensestraat 102, 3000 Leuven, Belgium. E-mail: hongwei.zhao@kuleuven.be

Abstract

In social sciences, researchers often compare relations between constructs, referred to as “structural relations”, across a large number of groups. This paper proposes Mixture Multigroup Bayesian SEM (MixMG-BSEM), a novel method for comparing structural relations across many groups while accounting for approximate measurement invariance in factor loadings. Traditional methods often assume exact measurement invariance, which may not reflect real-world data where small differences in measurement parameters commonly occur across many groups. MixMG-BSEM addresses this by using Multigroup Bayesian CFA with small-variance priors to allow for these small differences, and groups are then clustered based on their structural relations using Mixture Modeling. This is done in a stepwise estimation procedure built on the structural-after-measurement approach. By combining cluster-specific structural relations with small between-group differences in measurement parameters, MixMG-BSEM obtains a clustering that is driven only by the structural relations. The robustness and effectiveness of MixMG-BSEM are demonstrated through a simulation study.

Introduction

In social sciences, Structural Equation Modelling (SEM; Bollen, 1989; Hoyle, 2012) is widely used to investigate relations between constructs (e.g., emotions, motivation), referred to as “structural relations” within SEM. Researchers are often interested in how these structural relations vary across groups. For instance, Michael and Kyriakides (2023) examined how academic motivation mediated the effect of socioeconomic status on reading achievement among 15-year-old students and how this differed across 38 countries.

To study differences in structural relations, Multigroup SEM (MG-SEM) and Multilevel SEM (ML-SEM) can be used. MG-SEM estimates the structural relations for each group and allows testing whether they are equal across groups. ML-SEM captures variations in structural relations by normally distributed random effects around the overall mean estimate for each relation. Even though group-specific estimates of relations can be derived from random effects, only the mean and variance of each random effect are part of the model parameters, which makes ML-SEM more parsimonious, allowing for accurate parameter estimates in case of very small sample sizes per group. To pinpoint which groups have the same relations and for which groups they differ, MG-SEM and ML-SEM require pairwise comparisons of group-specific relations. As the number of groups increases, performing pairwise comparisons quickly becomes infeasible. For example, for 38 groups, this requires 703 pairwise comparisons per structural relation. To reduce the number of comparisons, mixture modeling (McLachlan & Peel, 2000) can be used to cluster groups based on similarity of the structural relations. Before performing such a clustering, it is essential to ensure that the structural relations are validly comparable across groups and that they are the only source of differences driving the clustering.

In social sciences, the constructs of interest are typically unobserved or latent variables, also known as “factors” in SEM. SEM addresses their latent nature by including a measurement model (MM), which specifies how latent variables are measured by observed indicators (often questionnaire items), whereas the relations of interest among the latent variables are part of the structural model (SM). For valid comparisons of constructs and their relations, measurement invariance (MI) must hold across the groups. MI implies that the MM is equal across groups, meaning that the constructs are measured in the same way, so that observed differences reflect differences in the constructs rather than differences in measurement.

MI is examined at different levels by assessing the equality of different subsets of MM parameters. Configural invariance evaluates whether the factor structure is the same across groups, meaning that, in each group, the same set of indicators relates to a factor. The strength and direction of the relations between factors and indicators are quantified by factor loadings. Whereas configural invariance only deals with which factor loadings are non-zero, weak or metric invariance requires the loadings to be equal across groups. Next, strong and strict invariance impose equality of the items' intercepts and residual or 'unique' variances, respectively. Metric invariance is a prerequisite for validly comparing structural relations (Davidov et al., 2012), whereas strong and strict invariance are not required. When full metric invariance (i.e., invariance of all loadings) does not hold, partial metric invariance (i.e., invariance of some loadings) still enables valid comparisons of structural relations (Byrne et al., 1989), as long as the loading differences are captured in the model (e.g., by group specific loadings). The same holds for differences in item intercepts and unique variances.

When combining SEM with mixture modeling, groups can be clustered on their structural relations by making the structural relations cluster-specific (i.e., the same for all groups assigned to a cluster). In traditional mixture SEM methods (Arminger & Stein, 1997; Dolan & van der Maas, 1998; Jedidi et al., 1997), MM parameters can be specified as invariant or cluster-specific, implying that MM differences can either be ignored or captured by the same clustering. To cluster groups only on the structural relations rather than also on differences in measurement, a framework of novel mixture SEM methods emerged recently. Perez Alonso and colleagues (2024) introduced Mixture Multigroup SEM (MixMG-SEM), which combines MG-SEM with mixture modeling. Zhao and colleagues (2024) proposed Mixture Multilevel SEM (MixML-SEM), which builds the mixture clustering onto the more parsimonious ML-SEM. The aim of both methods is to cluster groups specifically on the structural relations while accounting for measurement non-invariance, but the difference is that MixML-SEM uses Multilevel Confirmatory Factor Analysis (ML-CFA) with random effects to deal with MM differences, whereas MixMG-SEM uses Multigroup Confirmatory Factor Analysis (MG-CFA) with group-specific MM parameters. Their estimation builds on the stepwise "Structural-After-Measurement" (SAM; Rosseel & Loh, 2022) approach, where the MM is estimated first, using either MG-CFA or ML-CFA, followed by the SM, which includes clustering the groups on their structural relations. For comparability of the structural relations, both methods require at least partial metric invariance and impose exact equality for the invariant factor loadings (i.e., exact MI). However, with a large number of groups, achieving exact MI is often unrealistic. To address this, Multigroup Bayesian SEM (MG-BSEM; Muthén & Asparouhov, 2012, 2013) with Approximate MI (AMI) uses priors with small variances for the MM parameters to allow for small differences across groups while keeping them approximately equal. In this paper, we present Mixture Multigroup BSEM (MixMG-BSEM), which extends MG-BSEM with mixture modeling to cluster groups on the structural relations while capturing approximate invariance of factor loadings.

MixMG-BSEM, MixMG-SEM and MixML-SEM differ in their first estimation step only, that is, in their MM and the corresponding MI assumptions. MixMG-SEM and MixML-SEM require exact invariance for (at least) some loadings, whereas the first step of MixMG-BSEM is a MG-CFA with Bayesian estimation (MG-BCFA) that assumes approximately invariant loadings. Approximate invariance lies between exact invariance (where parameters are exactly equal across groups) and non-invariance (where parameters can differ substantially across groups), where exact invariance is more closely approximated as the variances of the priors become smaller.

The paper is structured as follows: We begin with a description of MixMG-BSEM in the Method section. Next, we evaluate its performance through a Simulation Study. Finally, the Discussion section summarizes the main findings and addresses limitations and future directions.

Method

As mentioned above, MixMG-BSEM is estimated in a stepwise manner, building on the SAM approach. In Step 1, MG-BCFA with small-variance priors is performed for each factor, and factor scores are extracted. In Step 2, these factor scores are used as single indicators to obtain group-specific factor covariances with Croon's correction (Croon, 2002). In Step 3, the SM is estimated, including the clustering and the cluster-specific structural relations, using an Expectation-Maximization (Dempster et al., 1977) algorithm for maximum likelihood estimation. Note that Steps 2 and 3 are the same as for MixML-SEM and are therefore only briefly described below (for details, see Zhao et al., 2024).

Step 1: Measurement Model with Bayesian Approximate Measurement Invariance

The MM defines how the factors are measured by the items and MG-CFA is used to compare MMs across groups. Note that we estimate the MM per factor, which implies that we assume the factors to be independent in Step 1. Indicating an individual in group g $(g = 1, \dots, G)$ by n_g and gathering the responses on the J_q items measuring factor q $(q = 1, \dots, Q)$ in the vector $\mathbf{x}_{n_g}$, the MG-CFA model for factor q is expressed as:

$$\mathbf{x}_{n_g} = \boldsymbol{\tau}_g + \boldsymbol{\lambda}_g \eta_{n_g} + \boldsymbol{\epsilon}_{n_g} \; with \; \boldsymbol{\epsilon}_{n_g} \sim MVN\left(\mathbf{0}, \boldsymbol{\Theta}_g\right) \tag{1}$$

where $\boldsymbol{\tau}_g$ is a J_q-dimensional vector of intercepts for group g, $\boldsymbol{\lambda}_g$ is a J_q-dimensional vector of factor loadings (i.e., item-factor relations) for group g, η_{n_g} denotes the latent variable score for the individual, and $\boldsymbol{\epsilon}_{n_g}$ is a J_q-dimensional vector of residuals, with the diagonal of $\boldsymbol{\Theta}_g$ containing the group-specific unique variances of the items. To set the scale of each factor, one can either set its variance to one or use the marker variable approach by fixing one loading (ideally, a strong and invariant loading) to one, for each group. In this paper, we adopt the marker variable approach to ensure that a one-unit change in the underlying factor has the same meaning across groups.

Since small differences in MM parameters are common across many groups and still allow for latent variable comparisons, we apply the assumption of approximate metric invariance (i.e., $\boldsymbol{\lambda}_g \approx \boldsymbol{\lambda}$ for all groups g) instead of exact invariance (i.e., $\boldsymbol{\lambda}_g = \boldsymbol{\lambda}$) in Step 1 of MixMG-BSEM. This is accomplished by using MG-CFA with Bayesian estimation[1] (MG-BCFA) and applying small-variance, normally distributed priors to the corresponding parameters, which constrain the group-specific parameters to be approximately equal. For this, both *Mplus* (Muthén & Muthén, 1998–2017) and the R-package *blavaan* (Merkle et al., 2021) are available, but we use *blavaan* by default because it is free and open-source. In *blavaan*, AMI is achieved by applying small-variance priors in every group except for the reference group, which is the first group (by default). A non-informative prior is used for the parameter in the first group and the parameter estimate for this group is used as the mean of the small-variance priors for that same parameter in the other groups.

[1] The possibility to use a different estimator in each step, such as Bayesian estimation for the MM and maximum likelihood for the SM (see also Zhao et al., 2024) is an important advantage of the SAM approach.

Since Bayesian estimation can be computationally challenging, two measures are taken to lower the computation time of Step 1 of MixMG-BSEM: (1) the data are centered per group to remove the mean structure (i.e., $\boldsymbol{\tau}_g = \mathbf{0}$ and $\boldsymbol{\alpha}_g = 0$), which is irrelevant to the comparison of structural relations, and (2) MG-BCFA is performed for each factor separately, which is in line with the "measurement blocks" approach in SAM (Rosseel & Loh, 2022) with one factor per measurement block. This approach lowers the number of parameters to be estimated and also enhances the model's robustness against MM misspecifications, such as unmodeled crossloadings.

In this paper, we assume that all factor loadings, except for the marker variable loadings, are approximately invariant, while the unique variances and factor variances are estimated as group-specific parameters (i.e., with non-informative priors per group). Note that it is harmless to specify exactly invariant loadings as approximately invariant since they will then be estimated as nearly identical across groups. Of course, in practice, combinations of exactly and approximately invariant loadings can be applied in Step 1 of MixMG-BSEM. Moreover, in theory, all combinations of exact invariance, approximate invariance and non-invariance can be used, but complex combinations may cause convergence problems.

To determine which parameters are (approximately) invariant or non-invariant, MI testing should be performed prior to using MixMG-BSEM. Note that, if exact invariance does not hold for a parameter, standard MG-CFA requires a tedious process of comparing group-specific parameter estimates to determine whether differences reflect non-invariance or approximate invariance. Instead, MG-BCFA (Muchen & Asparouhov, 2012) allows to test the tenability of AMI directly by imposing small-variance priors on MM parameters and assessing model fit. Muthén and Asparouhov (2012) recommend starting with a very small variance (e.g., 0.001) and, if needed, the priors' variances can be increased to reach a good model fit. In this way, MG-BCFA provides information on how large the parameter differences are (i.e., on the level of AMI). Model fit can be assessed using the posterior predictive *p* value (Gelman et al., 1996), but it is not very sensitive to the prior variances in case of large samples. Other fit measures include the Bayesian RMSEA (BRMSEA; Hoofs et al., 2018) and the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). The DIC balances model fit (i.e., the posterior mean deviance) and complexity (i.e., the effective number of parameters) in Bayesian models, with smaller values indicating a better balance. Regarding the prior selection in MG-BSEM, Kim et al. (2017) found that the DIC often selected models with smaller prior variances when the sample size was small and Pokropek et al. (2020) found that the DIC performed better as sample size increased, and recommended using the DIC with thresholds tailored to different sample sizes.

Once the marker variables and the approximately invariant loadings are confirmed by the MI testing, we obtain the specification of the MG-BCFA model that corresponds to the first step of MixMG-BSEM. In the next step, we need estimates of the factor scores and their uncertainty. To this end, the means and standard deviations of the posterior distributions of the individuals' factor scores (i.e., estimated latent variable scores) are appended to the data file.

Step 2: Single-Indicator Approach to Obtain Group-specific Factor Covariances

In a single-indicator approach, the factor scores are used as the "observed" proxy (or a single indicator) for the latent variable. Since factor scores are only estimates of the true latent variable scores, we apply Croon's correction (2002) to the factor score covariances ($\mathrm{cov}(\mathbf{f}_g)$) to obtain unbiased estimates of the true latent variable covariances ($\mathrm{cov}(\boldsymbol{\eta}_g)$), here denoted as $\boldsymbol{\Phi}_g^{s2}$:

$$\boldsymbol{\Phi}_g^{s2} = \widehat{\boldsymbol{\Lambda}_g}^{-1}\left(\mathrm{cov}(\mathbf{f}_g) - \widehat{\boldsymbol{\Theta}_g}\right)\left(\widehat{\boldsymbol{\Lambda}_g}'\right)^{-1} \tag{2}$$

where $\widehat{\boldsymbol{\Lambda}_g}$ corresponds to the $Q \times Q$ diagonal matrix of group-specific factor loadings (reflecting the reliability of the factor scores) and $\widehat{\boldsymbol{\Theta}_g}$ is the $Q \times Q$ diagonal matrix of group-specific unique variances. These estimates correspond to the MM parameters of the single indicators of the factors (i.e., the factor scores) rather than the original, observed indicators. These MM parameters are derived from the posterior mean and standard deviation estimates for the factor scores, obtained from Step 1. For details, please refer to Equations (7-8) in the MixML-SEM paper (Zhao et al., 2024).

Step 3: Structural Model with Mixture Clustering of the Groups

In Step 3, MixMG-BSEM clusters the groups and estimates cluster-specific structural relations. The SM is thus conditional on the cluster membership, z_{gk}, which denotes whether group g belongs to cluster k. Whereas the true cluster membership is assumed to be either 1 or 0, its estimation, $\hat{z}_{gk}$, ranges from 0 to 1 and represents the probability of group g belonging to cluster k. The model-implied factor covariance matrix $\boldsymbol{\Phi}_{gk}$, given that $z_{gk} = 1$, is defined as:

$$\boldsymbol{\Phi}_{gk} = (\mathbf{I} - \mathbf{B}_k)^{-1}\boldsymbol{\Psi}_{gk}(\mathbf{I} - \mathbf{B}_k)^{-1'} \tag{3}$$

where $\mathbf{B}_k$ contains the cluster-specific regression coefficients between latent variables, and $\boldsymbol{\Psi}_{gk}$ is the residual factor covariance matrix, which is specified as group-and-cluster-specific to ensure that clustering is driven only by the regressions $\mathbf{B}_k$ (for details, see Perez Alonso et al., 2024). The SM is estimated with maximum likelihood estimation using $\boldsymbol{\Phi}_g^{s2}$ as input.

For the mixture clustering in MixMG-BSEM, it is assumed that the (true) latent variable scores $\boldsymbol{\eta}_{n_g}$ are sampled from a mixture of K multivariate normal distributions. Specifically, all latent variable scores of group g, $\mathbf{H}_g$, are assumed to be sampled from the same distribution:

$$f(\mathbf{H}_g; \upsilon) = \sum_{k=1}^{K} \pi_k \prod_{n_g=1}^{N_g} MVN\left(\boldsymbol{\eta}_{n_g}; \boldsymbol{\alpha}_g, \boldsymbol{\Phi}_{gk}\right)\ with \sum_{k=1}^{K} \pi_k = 1 \tag{4}$$

where f is the population density function, υ represents the set of population parameters, and π_k is the prior probability that group g belongs to cluster k. The scores in $\mathbf{H}_g$are assumed to follow a normal distribution with $\boldsymbol{\alpha}_g$ as the factor mean (which is zero due to centering) and $\boldsymbol{\Phi}_{gk}$ as the factor covariance matrix. The unknown parameters υ are estimated by maximizing the following log-likelihood function:

$$\log L_\eta = \log\left(\prod_{g=1}^{G}\sum_{k=1}^{K}\pi_k \frac{1}{(2\pi)^{Q/2}|(\boldsymbol{\Phi}_{gk})|^{1/2}}\exp\left(-\frac{1}{2}tr\left(\boldsymbol{\Phi}_g^{s2}\boldsymbol{\Phi}_{gk}{}^{-1}\right)\right)^{N_g}\right)$$
$$= \sum_{g=1}^{G}\log\left(\sum_{k=1}^{K}\pi_k \frac{1}{(2\pi)^{Q/2}|(\boldsymbol{\Phi}_{gk})|^{1/2}}\exp\left(-\frac{1}{2}tr\left(\boldsymbol{\Phi}_g^{s2}\boldsymbol{\Phi}_{gk}{}^{-1}\right)\right)^{N_g}\right) \quad (5)$$

where $\boldsymbol{\Phi}_g^{s2}$ is the group-specific factor covariance matrix from Step 2 (Equation (2)), and $\boldsymbol{\Phi}_{gk}$ is the group-and-cluster-specific factor covariance matrix from Step 3 (Equation (3)). The maximum likelihood estimation is performed using the EM algorithm (Dempster et al., 1977). Specifically, in the E-step, the algorithm estimates the classification probabilities $\hat{z}_{gk}$ given the current parameter estimates. In the M-step, the algorithm estimates the unknown parameters υ given the classification probabilities obtained from the E-step. The E- and M-steps are iterated until convergence. A multi-start procedure is applied to mitigate convergence to local maxima, where the converged solution with the highest loglikelihood across the different starts is selected as the final result. For an in-depth explanation of the technical details of Step 3, readers are referred to Appendix A of the paper by Perez Alonso et al. (2024).

Simulation

In the simulation study, we evaluated the performance of MixMG-BSEM, assuming the true number of clusters was known. Firstly, we aimed to examine how MixMG-BSEM's performance was affected by factors related to the sample size, the number of clusters, the cluster sizes, the AMI of the loadings, and the size of (differences in) regression parameters. On top of that, since the first step of MixMG-BSEM estimates the MM per factor, we evaluated the consequences of ignoring crossloadings in this step. Literature on traditional SEM has shown that factor correlations tend to be overestimated when crossloadings are constrained to zero (e.g., Asparouhov et al., 2015; Marsh et al., 2009, 2010, 2014), which may affect the comparison of structural relations. However, given its stepwise estimation and measurement block approach, MixMG-BSEM may be relatively robust to overlooked crossloadings (Rosseel & Loh, 2022), but the recovery of clusters and regression parameters may still decline in case of multiple crossloadings. Secondly, we examined the impact of a key aspect of the Bayesian estimation; that is, the impact of different prior variances for the loadings on the recovery of clusters and cluster-specific regressions. We expected that using too narrow priors might fail to capture the loading differences across groups, which may affect the estimation of and clustering on the structural relations. Additionally, we also evaluated which prior was selected by the DIC, since selecting this prior is an important step in empirical practice.

In a complete factorial design, the following factors were manipulated:

1. Number of groups G (3 levels): 12, 24, 48;

2. Within-group sample size N_g (3 levels): 50, 100, 200;

3. Number of clusters K (2 levels): 2, 4;

4. Cluster sizes (2 levels): balanced, unbalanced;

5. Size of regression parameters β (2 levels): 0.2, 0.4;

6. Level of AMI for loadings (5 levels): 0.001, 0.005, 0.01, 0.05, 0.1;

7. Size of crossloadings (3 levels): 0, 0.2, 0,4

We chose a minimum of 12 groups with group sizes ranging from 50 to 200, which partially correspond to the group sizes in other simulation studies on Bayesian AMI (Kim et al., 2017; Lek et al. 2018). The number of groups in each cluster depended on the number of groups G, the number of clusters K and the cluster sizes. For the cluster sizes, in the balanced conditions, each cluster contained an equal number of groups. In the unbalanced conditions, the large cluster was three times the size of the small cluster, with the large cluster being randomly selected. For example, when $G = 24$ and $K = 4$, in unbalanced conditions, the large cluster contained 12 groups, and the remaining three clusters each contained four groups. Note that larger G, larger N_g, smaller K, and balanced cluster sizes result in larger within-cluster sample sizes, which were expected to improve the performance of MixMG-BSEM.

The data were generated from a SEM model with four latent variables, each measured by five items (see Fig 1), as in Perez Alonso et al. (2024) and Zhao et al. (2024). Specifically, the data were generated from a multivariate normal distribution (MVN) with covariance matrix $\mathbf{\Sigma}_{gk}$, determined by the parameters $\mathbf{B}_k$, $\mathbf{\Psi}_{gk}$, $\mathbf{\Lambda}_g$ and $\mathbf{\Theta}_{\boldsymbol{g}}$ (see Equation (6) in Perez Alonso et al., 2024).

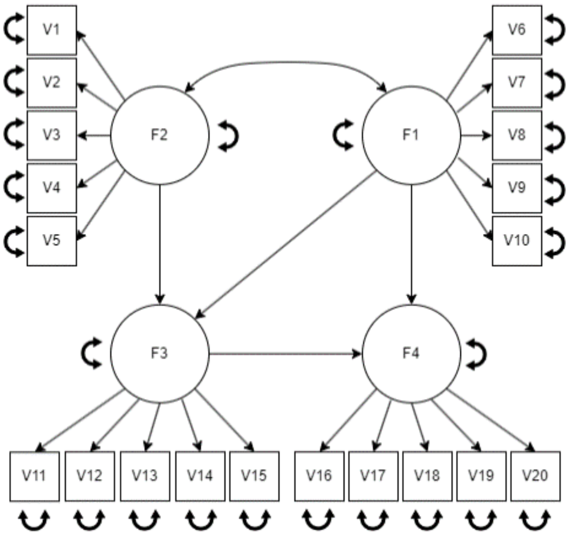


Fig 1. The data-generating model with exogenous factors F1 and F2 and endogenous factors F3 and F4.

The size of the regression parameters was set to β and, as shown in Fig 2, the differences between clusters were introduced by setting one regression parameter to zero in each cluster. Hence, larger values of β resulted in larger differences and thus in greater separation between clusters, which should make the clusters easier to recover.

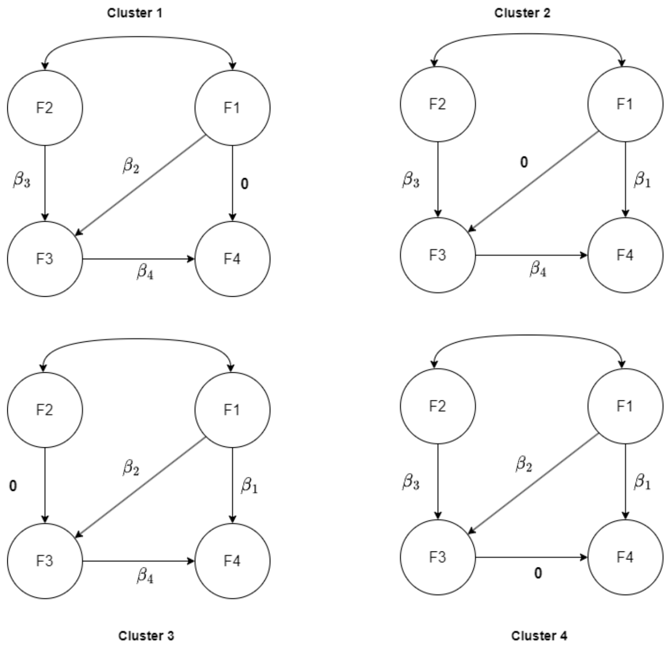


Fig 2. The cluster-specific structural relations.

For the group-and-cluster-specific residual factor covariances $\mathbf{\Psi}_{gk}$, we sampled the variances of the exogenous factors $F1$ and $F2$ from a uniform distribution $U(0.75, 1.25)$ and their covariance from $U(-0.3, 0.3)$. The total variances of the endogenous factors $F3$ and $F4$ were also sampled from $U(0.75, 1.25)$ and their residual variances are determined as follows: For $F3$ and $F4$, it was computed as $\mathrm{Var}(F3)_g - (\beta_{2,k}^2\mathrm{Var}(F1)_g + \beta_{3,k}^2\mathrm{Var}(F2)_g + 2\beta_{2,k}\beta_{3,k}\mathrm{Cov}(F1, F2)_g$ and $\mathrm{Var}(F4)_g - (\beta_{1,k}^2\mathrm{Var}(F1)_g + \beta_{4,k}^2\mathrm{Var}(F3)_g + 2\beta_{1,k}\beta_{4,k}\left(\beta_{2,k}\mathrm{Var}(F1)_g + \beta_{3,k}\mathrm{Cov}(F1, F2)_g\right))$, respectively.

In loading matrix $\mathbf{\Lambda}_g$, the first loading of each factor was fixed to one. The other loadings (except for crossloadings) were approximately invariant across groups and were sampled from a normal distribution with a mean of $\sqrt{0.4}$ and a variance that depended on the level of the AMI.

For instance, to obtain an AMI level of 0.01, which implies a variance of 0.01 for differences in loadings, we sampled loadings from a normal distribution with a variance of 0.005 for all groups.[2] Per factor, one crossloading was added to the third item measuring the next factor (i.e., item 8 crossloaded on factor 1, item 13 on factor 2, item 18 on factor 3, and item 3 on factor 4). A value of 0 corresponded to no crossloading, 0.2 to a moderate crossloading, and 0.4 to a large crossloading. The unique variances on the diagonal of $\mathbf{\Theta}_g$ were sampled from $U(0.50, 0.70)$.

Finally, the data were sampled from $\text{MVN}(\mathbf{0}, \mathbf{\Sigma}_{gk})$ for each group. In total, we generated 3 (number of groups) × 3 (within-group sample size) × 2 (number of clusters) × 2 (cluster sizes) × 2 (size of regression parameters) × 5 (size of AMI) × 3 (size of crossloadings) × 50 (replications) = 54,000 data sets according to the described procedure, using R version 4.2.1 (R Core Team, 2022). All data sets were analyzed with MixMG-BSEM with 50 random starts and the true number of clusters. For each data set, we performed the analysis five times, with different prior variances for the loadings (i.e., 0.001, 0.005, 0.01, 0.05, 0.1) in Step 1, to examine the performance of MixMG-BSEM across different prior variances. The average computation time was 35.9 minutes ($SD = 22.6$) for Step 1 (mainly influenced by G and N_g), 0.02 minutes ($SD = 0.02$) for the intermediate Step 2, and 2.8 minutes ($SD = 4.2$) for Step 3 (mainly influenced by N_g and β). [3]

[2] In *blavaan*, the estimate of a parameter in the first group is used as the mean of the prior for that same parameter in the other groups. Consequently, the prior reflects the differences of the other groups to the reference group. The variance of the difference between two factor loadings equals the sum of their individual variances, assuming there is no covariance between them. For all groups, including the reference group, we sampled loadings from a normal distribution with a variance that is half the targeted MI level for all groups, so that the variance of the loading differences toward the reference group equals the targeted MI level.

[3] The first step of MixMG-BSEM (i.e., estimating the MM using *blavaan*) can be computationally demanding, especially for larger sample sizes. Luckily, the stepwise estimation of MixMG-BSEM implies that the MM needs to be estimated only once, even when estimating the SM with different numbers of clusters for model selection. Alternatively, *Mplus* offers a more time-efficient estimation of the MM with AMI, though it is commercial software. For instance, for a dataset with 48 groups and 200 observations per group, *blavaan* took around 118 minutes (without parallelization), while *Mplus* took only 3 minutes. Eliminating the mean structure by centering per group (see Method section) clearly helped, since the computation times of *blavaan* and *Mplus* increased to 152 and 122 minutes, respectively, when including the mean structure in the model.

Results

Recovery of factor loadings.

We evaluated the recovery of the group-specific factor loading estimates for each item j, using the mean error (ME) and the Root Mean Squared Error (RMSE) across groups as follows:

$$ME_{\lambda_j} = \frac{\sum_{g=1}^{G}\left(\hat{\lambda}_{gj} - \lambda_{gj}\right)}{G} \tag{6}$$

$$RMSE_{\lambda_j} = \sqrt{\frac{\sum_{g=1}^{G}\left(\hat{\lambda}_{gj} - \lambda_{gj}\right)^2}{G}} \tag{7}$$

where λ_{gj} is the true group-specific loading of the j-th item on the factor, and $\hat{\lambda}_{gj}$ is the corresponding estimate. For items with crossloadings, we expect the loadings to be overestimated in all groups when the crossloadings are ignored, resulting in a positive ME_{λ_j}. Note that when averaging ME_{λ_j} across replications, for instance, across all datasets pertaining to a certain level of G (Table 1), the result is equivalent to a measure of bias (i.e., the difference between the estimated and true loading values for each λ_{gj}, averaged across replications), averaged across the groups.

When using MixMG-BSEM with the true prior variances for the loadings, the average ME_{λ_j}across the four factors and all simulated data sets was 0.010, 0.051, 0.010, and 0.010, respectively, for the loadings of the second to the fifth item of each factor (Table 1, last row). Note that ME_{λ_3} was larger due to the disregarded crossloadings on that item. This was also the only ME_{λ_j} value that differed across the four factors. Specifically, the ME_{λ_3} values were 0.059, 0.010, 0.064, and 0.069 for $F1$ to $F4$, respectively. It seems that the third loading for $F2$ is unaffected by the ignored crossloading, which may be explained by the fact that, unlike the other factors, $F2$ is involved in only one direct regression relation with the other factors[4] and is thus less correlated with the other factors. In conditions without crossloadings, ME_{λ_3} is the same across all factors, with a value of 0.11. The average $RMSE_{\lambda_j}$ was 0.039, 0.075, 0.039, and 0.039, respectively (Table 2, last row), where only $RMSE_{\lambda_3}$ differed across factors (i.e., 0.076, 0.047, 0.081, and 0.086 for $F1$ to $F4$, respectively). When the crossloadings were zero (i.e., without crossloadings), ME_{λ_3} and $RMSE_{\lambda_3}$ took on similar values as for the other loadings ($ME_{\lambda_3} = 0.011$ and $RMSE_{\lambda_3} = 0.039$), whereas they increased with larger crossloadings: with crossloadings of 0.2, $ME_{\lambda_3} = 0.051$ and $RMSE_{\lambda_3} = 0.071$; and with crossloadings of 0.4, $ME_{\lambda_3} = 0.091$ and $RMSE_{\lambda_3} = 0.116$. For the third loading, ME and RMSE were also higher in case of larger regression coefficients (β), which imply stronger correlations between factors. Specifically, with $\beta = 0.2$, $ME_{\lambda_3} = 0.036$ and $RMSE_{\lambda_3} = 0.060$; and with $\beta = 0.4$, $ME_{\lambda_3} = 0.065$ and $RMSE_{\lambda_3} = 0.090$. Note that larger N_g and smaller levels of AMI – thus applying lower prior variances – resulted in lower ME and RMSE values for all items. The latter is explained by the fact that a lower prior variance more strongly approximates an equality constraint, which lowers the sample size requirements.

Table 1. The average ME_{λ_j} (standard deviation, SD, in brackets) for factor loading estimates when using the true prior variances for the loadings.

Factor	Level	ME_{λ_2}	ME_{λ_3}	ME_{λ_4}	ME_{λ_5}
G	12	0.011 (0.011)	0.051 (0.040)	0.011 (0.011)	0.011 (0.011)
	24	0.010 (0.011)	0.050 (0.039)	0.010 (0.011)	0.010 (0.011)
	48	0.010 (0.011)	0.050 (0.039)	0.010 (0.011)	0.010 (0.011)
K	2	0.010 (0.011)	0.051 (0.040)	0.010 (0.011)	0.010 (0.011)
	4	0.010 (0.011)	0.050 (0.039)	0.010 (0.011)	0.010 (0.011)
Cluster sizes	balanced	0.010 (0.011)	0.051 (0.039)	0.010 (0.011)	0.010 (0.011)

[4] It has indirect relations with the other factors via the correlation between F1 and F2, but the expected value of this correlation is zero.

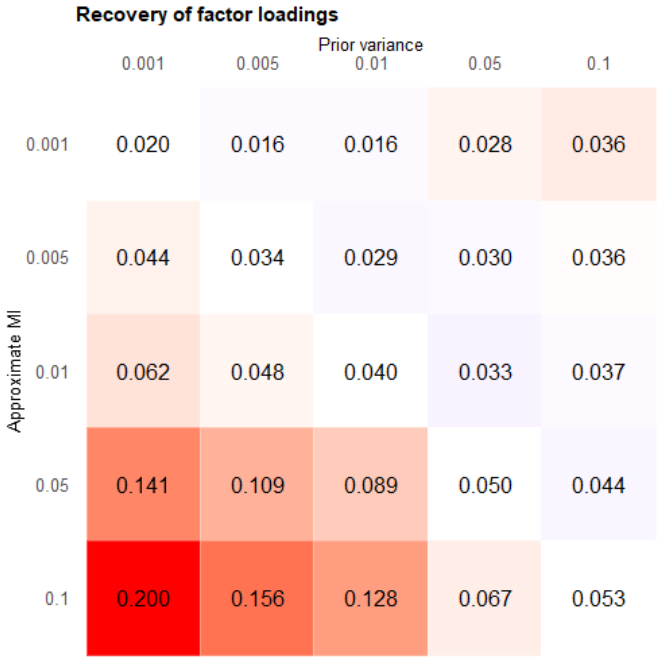
Factor	Level	ME_{λ_2}	ME_{λ_3}	ME_{λ_4}	ME_{λ_5}
	unbalanced	0.010 (0.011)	0.051 (0.040)	0.010 (0.011)	0.010 (0.011)
N_g	50	0.015 (0.016)	0.056 (0.041)	0.015 (0.016)	0.015 (0.016)
	100	0.010 (0.008)	0.050 (0.039)	0.010 (0.008)	0.010 (0.008)
	200	0.006 (0.004)	0.046 (0.038)	0.006 (0.004)	0.006 (0.004)
β	0.2	0.011 (0.011)	0.036 (0.024)	0.011 (0.011)	0.011 (0.011)
	0.4	0.010 (0.011)	0.065 (0.046)	0.010 (0.011)	0.010 (0.011)
AMI	0.001	-0.001 (0.002)	0.038 (0.037)	-0.001 (0.002)	-0.001 (0.002)
	0.005	0.004 (0.002)	0.044 (0.037)	0.004 (0.001)	0.004 (0.001)
	0.01	0.008 (0.002)	0.048 (0.038)	0.008 (0.002)	0.008 (0.002)
	0.05	0.019 (0.008)	0.059 (0.039)	0.019 (0.008)	0.019 (0.008)
	0.1	0.022 (0.012)	0.063 (0.040)	0.022 (0.012)	0.022 (0.012)
Crossloadings	0	0.011 (0.011)	0.011 (0.011)	0.011 (0.011)	0.011 (0.011)
	0.2	0.010 (0.011)	0.051 (0.019)	0.010 (0.011)	0.010 (0.011)
	0.4	0.010 (0.011)	0.091 (0.032)	0.010 (0.011)	0.010 (0.011)
Total		0.010 (0.011)	0.051 (0.040)	0.010 (0.011)	0.010 (0.011)

Table 2. The average $RMSE_{\lambda_j}$ (SD in brackets) for factor loading estimates when using the true prior variances for the loadings.

Factor	Level	$RMSE_{\lambda_2}$	$RMSE_{\lambda_3}$	$RMSE_{\lambda_4}$	$RMSE_{\lambda_5}$
G	12	0.040 (0.019)	0.076 (0.042)	0.040 (0.019)	0.040 (0.019)
	24	0.039 (0.017)	0.075 (0.041)	0.039 (0.017)	0.039 (0.017)
	48	0.038 (0.016)	0.074 (0.041)	0.038 (0.016)	0.038 (0.016)
K	2	0.039 (0.017)	0.076 (0.041)	0.039 (0.017)	0.039 (0.017)
	4	0.039 (0.017)	0.075 (0.041)	0.039 (0.017)	0.039 (0.017)
Cluster sizes	balanced	0.039 (0.017)	0.075 (0.041)	0.039 (0.017)	0.039 (0.017)
	unbalanced	0.039 (0.017)	0.076 (0.041)	0.039 (0.017)	0.039 (0.017)
N_g	50	0.051 (0.021)	0.085 (0.042)	0.051 (0.021)	0.051 (0.021)
	100	0.038 (0.012)	0.075 (0.039)	0.038 (0.012)	0.038 (0.012)
	200	0.028 (0.007)	0.067 (0.040)	0.028 (0.007)	0.028 (0.007)

Factor	Level	$RMSE_{\lambda_2}$	$RMSE_{\lambda_3}$	$RMSE_{\lambda_4}$	$RMSE_{\lambda_5}$
β	0.2	0.039 (0.017)	0.060 (0.026)	0.039 (0.017)	0.039 (0.017)
	0.4	0.039 (0.017)	0.090 (0.048)	0.039 (0.017)	0.039 (0.017)
AMI	0.001	0.020 (0.002)	0.055 (0.037)	0.020 (0.002)	0.020 (0.002)
	0.005	0.034 (0.006)	0.067 (0.035)	0.034 (0.006)	0.034 (0.006)
	0.01	0.040 (0.009)	0.074 (0.036)	0.040 (0.009)	0.040 (0.009)
	0.05	0.050 (0.017)	0.088 (0.041)	0.050 (0.017)	0.050 (0.017)
	0.1	0.053 (0.020)	0.093 (0.044)	0.053 (0.020)	0.053 (0.020)
Crossloadings	0	0.039 (0.017)	0.039 (0.017)	0.039 (0.017)	0.039 (0.017)
	0.2	0.039 (0.017)	0.071 (0.023)	0.039 (0.017)	0.039 (0.017)
	0.4	0.039 (0.017)	0.116 (0.036)	0.039 (0.017)	0.039 (0.017)
Total		0.039 (0.017)	0.075 (0.041)	0.039 (0.017)	0.039 (0.017)

To illustrate the effect of the prior variances for the loadings, $RMSE_{\lambda_2}$ across different prior variances is shown in Fig 3. The diagonal of the plot represents cases where the prior variances were correctly specified, while the lower part shows cases where the priors were narrower than the true level of AMI. Overall, we see that applying a too narrow prior resulted in a larger $RMSE_{\lambda_2}$. In general, applying the true priors or slightly wider priors resulted in lower $RMSE_{\lambda_2}$ values. Perhaps, a slightly wider prior allowed to capture some additional loading differences due to sampling fluctuations.



Approximate MI	0.001	0.005	0.01	0.05	0.1
0.001	0.020	0.016	0.016	0.028	0.036
0.005	0.044	0.034	0.029	0.030	0.036
0.01	0.062	0.048	0.040	0.033	0.037
0.05	0.141	0.109	0.089	0.050	0.044
0.1	0.200	0.156	0.128	0.067	0.053

Fig 3. $RMSE_{\lambda_2}$ across different prior variances, indicated by the columns, whereas the rows represent the true levels of AMI. The diagonal contains cases where the prior variances were correctly specified, while the lower part represents cases where the priors were too narrow. For each row, the cells are colored in red if the $RMSE_{\lambda_2}$ is larger than the $RMSE_{\lambda_2}$ on the diagonal, and in blue if it is smaller.

Since the prior variance affects the loading recovery, we also evaluated prior selection using the DIC. When looking at the prior selection per loading, the correct selection rate was 28.7% across all loadings and all simulated data sets.[5] For 21.4% of the data sets, the prior selection was flawless in the sense that the true priors were selected for *all* loadings. Generally, the DIC tended to select smaller prior variances. Specifically, for AMI levels of 0.001, 0.005, and 0.01, DIC most often selected a prior variance of 0.001, with selection rates of 100%, 100%, and 98.7%, respectively, averaged across loadings. For an AMI level of 0.05, DIC primarily selected a prior variance of 0.01 (52.0%), followed by prior variances of 0.05 (21.5%) and 0.005 (15.5%). For an AMI level of 0.1, DIC mostly selected prior variances of 0.05 (70.9%) and 0.1 (22.4%). Thus, overall, prior selection based on the DIC is not satisfactory, especially considering the larger ME and RMSE for loading estimates when using too narrow priors.

[5] Similar results were found with the widely applicable information criterion (WAIC; Watanabe, 2010) and leave-one-out information criterion (LOOIC; Geisser & Eddy, 1979; Gelfand & Dey, 1994): WAIC: 28.8%; LOOIC: 28.8%.

Sensitivity to local maxima.

To evaluate how often (Step 3 of) MixMG-BSEM converged to a local maximum, we compared the log-likelihood of the final best solution (out of 50 random starts) to the one obtained when starting from the true clustering, which is a proxy for the global maximum. If $\log L_\eta$ was more than 0.001 lower than the proxy, the solution was considered a local maximum. Overall, when applying the true priors, MixMG-BSEM ended up in a local maximum for 1.81% of the data sets, with all local maxima occurring in case of unbalanced cluster sizes.

Recovery of clusters.

The Adjusted Rand Index (ARI; Hubert & Arabie, 1985) measures the similarity between two partitions while correcting for chance, with a value of one indicating perfect agreement and zero indicating the level of agreement between two random partitions. To compute the ARI, the modal clustering (i.e., assigning each group to the cluster with the highest classification probability) was compared to the true clustering. Additionally, the correct clustering rate (%CC) was computed based on an indicator variable that equals 1 for a perfect cluster recovery (i.e., ARI = 1), and 0 otherwise.

When using the true priors, the average ARI across all simulated data was 0.882 and the correct clustering rate was 87.6%. As expected, fewer clusters, balanced cluster sizes, larger groups, and larger regression coefficients contributed to better cluster recovery (Table 3). The cluster recovery was the worst when N_g was 50 ($ARI = 0.647$ and $\%CC = 63.3\%$), whereas increasing it to 100 significantly improved the recovery ($ARI = 0.999$ and $\%CC = 99.6\%$).

Across different prior variances, the ARI slightly increased with wider priors. For example, for an AMI level of 0.001, the ARI increased from 0.878 to 0.888 when a wider prior was applied (i.e., when the prior variance increased from 0.001 to 0.1). When the applied prior was too narrow, the ARI slightly dropped. For an AMI level of 0.1, it decreased from 0.891

when using the true prior variance to 0.875 when using a prior variance of 0.001. [6] This may be related to the worse loading recovery observed with too narrow priors (Fig 3).

Table 3. The average ARI and correct clustering rate (%CC) (SD in brackets) when using the true prior variances for the loadings.

Factor	Level	ARI	%CC
G	12	0.882 (0.318)	0.875 (0.330)
	24	0.883 (0.317)	0.877 (0.328)
	48	0.880 (0.322)	0.876 (0.330)
K	2	0.930 (0.248)	0.922 (0.268)
	4	0.834 (0.371)	0.830 (0.375)
Cluster sizes	balanced	0.916 (0.278)	0.915 (0.279)
	unbalanced	0.848 (0.353)	0.837 (0.369)
N_g	50	0.647 (0.471)	0.633 (0.482)
	100	0.999 (0.015)	0.996 (0.067)
	200	1.000 (0.003)	1.000 (0.013)
β	0.2	0.764 (0.419)	0.752 (0.432)
	0.4	1.000 (0.001)	1.000 (0.011)
AMI	0.001	0.878 (0.326)	0.875 (0.331)
	0.005	0.878 (0.326)	0.875 (0.331)
	0.01	0.879 (0.324)	0.875 (0.331)
	0.05	0.885 (0.314)	0.878 (0.327)
	0.1	0.891 (0.305)	0.878 (0.327)
Crossloadings	0	0.883 (0.319)	0.878 (0.327)
	0.2	0.881 (0.320)	0.876 (0.329)
	0.4	0.882 (0.318)	0.874 (0.332)
Total		0.882 (0.319)	0.876 (0.329)

[6] To evaluate the recovery of clusters with exact (rather than approximate) MI constraints on factor loadings, we ran MixMG-SEM (Perez Alonso et al., 2024) for the first 25 replications. The average ARI values were 0.877, 0.876, 0.875, 0.875, and 0.868 when the approximate AMI levels in the data-generating model were 0.001 to 0.1, respectively – all of which a bit lower than the ARI for MixMG-BSEM when using a prior variance of 0.001.

Recovery of regression parameters.

To evaluate the recovery of the regression parameters, we computed the $RMSE_\beta$ per regression parameter (i.e., β_1, β_2, β_3, and β_4):

$$RMSE_\beta = \sqrt{\frac{\sum_{k=1}^{K}\left(\hat{\beta}_k - \beta_k\right)^2}{K}} \tag{8}$$

where $\hat{\beta}_k$ is the estimated regression coefficient in cluster k and β_k is the corresponding true value. Note that the estimated regression coefficients can deviate from the true values in either direction, being over- or underestimated. When averaged across clusters, the deviations can thus cancel each other out which is why ME_β is not reported.

On average, $RMSE_\beta$ was 0.050, 0.022, 0.051, and 0.046 (Table 4) for β_1, β_2, β_3, and β_4, respectively. Similar to the trends observed for the cluster recovery, fewer clusters, balanced cluster sizes, larger groups, and larger regression coefficients resulted in smaller $RMSE_\beta$. Larger crossloadings resulted in larger $RMSE_\beta$ values, with β_2 being the least affected. This is expected as β_2 is the only regression parameter between factors not involving ignored crossloadings (i.e., no crossloadings between $F1$ and $F3$). Note that the recovery of the regression parameters was barely affected by using different prior variances, even more narrow ones, likely due to the fact that the cluster recovery was hardly affected as well.

Table 4. The average $RMSE_\beta$ (SD in brackets) for each of the four estimated regression parameters when using the true prior variances for the loadings.

Factor	Level	$RMSE_{\beta_1}$	$RMSE_{\beta_2}$	$RMSE_{\beta_3}$	$RMSE_{\beta_4}$
G	12	0.051 (0.034)	0.022 (0.027)	0.052 (0.036)	0.049 (0.028)
	24	0.050 (0.035)	0.022 (0.028)	0.051 (0.036)	0.046 (0.028)
	48	0.050 (0.035)	0.022 (0.029)	0.051 (0.036)	0.043 (0.028)

Factor	Level	$RMSE_{\beta_1}$	$RMSE_{\beta_2}$	$RMSE_{\beta_3}$	$RMSE_{\beta_4}$
K	2	0.047 (0.034)	0.017 (0.026)	0.046 (0.033)	0.040 (0.025)
	4	0.053 (0.035)	0.026 (0.029)	0.057 (0.038)	0.052 (0.030)
Cluster sizes	balanced	0.049 (0.032)	0.020 (0.023)	0.051 (0.036)	0.045 (0.028)
	unbalanced	0.052 (0.037)	0.024 (0.032)	0.052 (0.036)	0.047 (0.029)
N_g	50	0.068 (0.038)	0.045 (0.037)	0.064 (0.039)	0.057 (0.033)
	100	0.043 (0.029)	0.011 (0.008)	0.047 (0.033)	0.042 (0.024)
	200	0.041 (0.029)	0.010 (0.009)	0.044 (0.033)	0.040 (0.023)
β	0.2	0.060 (0.039)	0.030 (0.037)	0.056 (0.039)	0.052 (0.033)
	0.4	0.041 (0.027)	0.014 (0.009)	0.047 (0.033)	0.041 (0.021)
AMI	0.001	0.050 (0.035)	0.022 (0.029)	0.049 (0.036)	0.045 (0.028)
	0.005	0.051 (0.035)	0.022 (0.029)	0.050 (0.036)	0.045 (0.028)
	0.01	0.051 (0.035)	0.022 (0.028)	0.051 (0.036)	0.046 (0.028)
	0.05	0.051 (0.034)	0.022 (0.027)	0.053 (0.036)	0.047 (0.029)
	0.1	0.050 (0.034)	0.022 (0.026)	0.053 (0.036)	0.047 (0.029)
Crossloadings	0	0.019 (0.028)	0.017 (0.029)	0.013 (0.024)	0.023 (0.023)
	0.2	0.051 (0.022)	0.021 (0.028)	0.053 (0.016)	0.045 (0.019)
	0.4	0.082 (0.019)	0.028 (0.025)	0.089 (0.013)	0.070 (0.019)
Total		0.050 (0.035)	0.022 (0.028)	0.051 (0.036)	0.046 (0.028)

Conclusion

We assessed the performance of MixMG-BSEM when the true number of clusters is known. We found that performing 50 random starts in Step 3 largely prevented local maxima. The recovery of clusters and regression parameters was good to excellent when the within-group sample size was at least 100 and/or in case of a larger cluster separation (i.e., $\beta = 0.4$). Ignoring crossloadings (by estimating the MM per factor) resulted in biased estimates for factor loadings and regression parameters, but barely affected the clustering. DIC tended to select too narrow prior variances, which come with a worse recovery of factor loadings. Luckily, the recovery of clusters and regression parameters was relatively robust to using too narrow priors.

Discussion

We presented MixMG-BSEM as a new addition to the novel mixture SEM framework for comparing structural relations across many groups. Unlike the existing approaches that rely on the exact MI assumption, MixMG-BSEM adopts the more realistic assumption of AMI, which accommodates small differences in MM parameters across groups. Specifically, after estimating the MM using MG-BCFA with small-variance priors, MixMG-BSEM clusters groups with the same structural relations, thereby eliminating the need for pairwise comparisons of group-specific structural relations.

Currently, MixMG-BSEM estimates the MM per factor (i.e., with one factor per measurement block). In the simulation study, the cluster recovery was unaffected by ignoring crossloadings, but the recovery of the factor loadings and regression estimates was affected. Therefore, it would be valuable to investigate the performance of MixMG-BSEM when including factors with crossloadings in the same measurement block, at the cost of a longer computation time. In that case, small-variance priors could also be applied to the crossloadings to allow for small differences (Muthén & Asparouhov, 2012). However, it is important to note that the default prior mean for crossloadings is zero, whereas applying a prior mean of zero to a sizeable crossloading can negatively impact the regression parameter estimates (Wei et al., 2022). Therefore, researchers should gather prior information about crossloadings before choosing an appropriate prior (Wei et al., 2022).

While the simulation study evaluated the performance of MixMG-BSEM with approximate metric invariance for all loadings, except for the invariant marker variable loading, MixMG-BSEM can theoretically accommodate all combinations of exact, approximate and non-invariance for the loadings. The stepwise estimation of MixMG-BSEM conveniently allows to tweak the MG-BCFA model, for instance, by specifying certain loadings as non-invariant, before moving onto the next steps. Similarly, if group-specific loading estimates are virtually identical across groups, one may consider specifying the loading as exactly invariant. Specifying an invariant parameter as approximately invariant is rather harmless, whereas specifying a non-invariant parameter as approximately invariant may introduce bias in parameter estimation and affect the clustering. Note that MG-BCFA allows to evaluate non-invariance for all parameters, which is achieved by comparing group-specific estimates to the credible intervals of the average posterior estimates across all groups (e.g., Winter & Depaoli, 2020). In future research, it would be interesting to evaluate the performance of MixMG-BSEM when non-invariant loadings are specified as approximately invariant.

The simulation study assumed the number of clusters to be known, whereas this is typically unknown for empirical data. To determine the number of clusters, different methods are available, such as the Bayesian Information Criterion (BIC; Schwarz, 1978), Akaike Information Criterion (AIC; Akaike, 1973), and convex hull procedure (CHull; Ceulemans & Kiers, 2006). In brief, all these methods balance model fit (i.e., the log-likelihood) and model complexity (i.e., the number of parameters). BIC and AIC do so by combining model fit and a penalty for model complexity into a single criterion, whereas CHull uses a generalized scree test. Previous studies on model selection for MixMG-SEM (Perez Alonso et al., 2024) and MixML-SEM (Zhao et al., 2024) have shown that combining AIC and CHull – with visual inspection of the scree plot – is an effective way to determine the number of clusters. Since MixMG-BSEM performs the same mixture clustering on group-specific factor covariances as these methods, we expect these recommendations to generalize to MixMG-BSEM. However, in the future, it would still be useful to evaluate model selection for MixMG-BSEM specifically.

Currently, MixMG-BSEM combines Bayesian and maximum likelihood estimation, assuming continuous items. In empirical practice, we often work with ordinal items with a few response categories (e.g., Likert scale items). To accommodate ordinal data in MixMG-BSEM, only the first step (i.e., MG-BCFA) would need to be adjusted to deal with ordinal data (Muthén & Asparouhov, 2013), whereas the subsequent steps would remain unchanged. In future studies, it will be valuable to evaluate the performance of MixMG-BSEM adapted to ordinal data.

In conclusion, MixMG-BSEM is an effective method for accommodating AMI while clustering structural relations of interest. By relaxing the strict assumption of exact MI, it extends the framework of novel mixture SEM methods in an important way, making it more suited for empirical applications where small differences in parameters across groups are expected.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In B. N. Petrov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). https://doi.org/10.1007/978-1-4612-1694-0_15

Arminger, G., & Stein, P. (1997). Finite Mixtures of Covariance Structure Models with Regressors: Loglikelihood Function, Minimum Distance Estimation, Fit Indices, and a Complex Example. *Sociological Methods & Research*, *26*(2), 148–182. https://doi.org/10.1177/0049124197026002002

Asparouhov, T., Muthén, B., & Morin, A. J. S. (2015). Bayesian Structural Equation Modeling with Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. *Journal of Management*, *41*(6), 1561–1577. https://doi.org/10.1177/0149206315591075

Bollen, K. A. (1989). Structural equations with latent variables (pp. xiv, 514). John Wiley & Sons. https://doi.org/10.1002/9781118619179

Byrne, B. M., Shavelson, R. J., & Muthén, B. (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

Croon, M. A. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent Variable and Latent Structure Models* (pp. 195–224). Lawrence Erlbaum.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a Multilevel Structural Equation Modeling Approach to Explain Cross-Cultural Measurement Noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. https://doi.org/10.1177/0022022112438397

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*(3), 227–253. https://doi.org/10.1007/BF02294853

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. *Statistica Sinica*, *6*(4), 733–760.

Hoofs, H., Van De Schoot, R., Jansen, N. W. H., & Kant, Ij. (2018). Evaluating Model Fit in Bayesian Confirmatory Factor Analysis with Large Samples: Simulation Study Introducing the BRMSEA. *Educational and Psychological Measurement*, *78*(4), 537–568. https://doi.org/10.1177/0013164417709314

Hoyle, R. H. (2012). Handbook of structural equation modeling. Guilford press.

Huang, S., & Yin, H. (2024). The relationships between paternalistic leadership, teachers' emotional labor, engagement, and turnover intention: A multilevel SEM analysis. *Teaching and Teacher Education*, *143*, 104552. https://doi.org/10.1016/j.tate.2024.104552

Jedidi, K., Jagpal, H., & Desarbo, W. (1997). Finite-Mixture Structural Equation Models for Response-Based Segmentation and Unobserved Heterogeneity. *Marketing Science*, *16*, 39–59. https://doi.org/10.1287/mksc.16.1.39

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017a). 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

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017b). 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

Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., De Roover, K., Jak, S., Meitinger, K., Menold, N., Muthén, B., Rudnev, M., Schmidt, P., & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, *110*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

Marsh, H. W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A. J. S., Trautwein, U., & Nagengast, B. (2010). A new look at the big five factor structure through exploratory structural equation modeling. *Psychological Assessment*, *22*(3), 471–491. https://doi.org/10.1037/a0019227

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439–476. https://doi.org/10.1080/10705510903008220

McLachlan, G., & Peel, D. (2000). Finite Mixture Models (1st ed.). Wiley. https://doi.org/10.1002/0471721182

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient Bayesian Structural Equation Modeling in *Stan*. *Journal of Statistical Software*, *100*(6). https://doi.org/10.18637/jss.v100.i06

Michael, D., & Kyriakides, L. (2023). Mediating effects of motivation and socioeconomic status on reading achievement: A secondary analysis of PISA 2018. *Large-Scale Assessments in Education*, *11*(1), 31. https://doi.org/10.1186/s40536-023-00181-9

Muthén, L. K., & Muthén, B. O. (1998-2017). Mplus user's guide (8th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2013). BSEM Measurement Invariance Analysis. *Mplus Web Notes: No. 17*. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B., & Asparouhov, T. (2013). New Methods for the Study of Measurement Invariance with Many Groups. Retrieved from https:// www.statmodel.com/download/PolAn.pdf

Perez Alonso, A. F., Rosseel, Y., Vermunt, J. K., & De Roover, K. (2024). Mixture multigroup structural equation modeling: A novel method for comparing structural relations across many groups. *Psychological Methods*. https://doi.org/10.1037/met0000667

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing Priors in Bayesian Measurement Invariance Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling:*

A Multidisciplinary Journal, *27*(5), 750–764. https://doi.org/10.1080/10705511.2019.1703708

R Core Team (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/met0000503

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Wei, X., Huang, J., Zhang, L., Pan, D., & Pan, J. (2022). Evaluation and Comparison of SEM, ESEM, and BSEM in Estimating Structural Models with Potentially Unknown Cross-loadings. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 327–338. https://doi.org/10.1080/10705511.2021.2006664

Winter, S. D., & Depaoli, S. (2020). An illustration of Bayesian approximate measurement invariance with longitudinal data and a small sample size. *International Journal of Behavioral Development*, *44*(4), 371–382. https://doi.org/10.1177/0165025419880610

Zhao, H., Vermunt, J. K., & De Roover, K. (2024). Mixture Multilevel SEM versus Multilevel SEM for comparing structural relations across groups in presence of measurement non-invariance. https://doi.org/10.31234/osf.io/cdrhv