METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Selecting the Number of Clusters in Mixture Multigroup Structural Equation Modeling

Andres F. Perez Alonso [1,2], Jeroen K. Vermunt [1], Yves Rosseel [3], Kim De Roover [1,2]

[1] *Department of Methodology and Statistics, Tilburg University, Tilburg, the Netherlands.* **[2]** *Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium.* **[3]** *Department of Data Analysis, Ghent University, Ghent, Belgium.*

Methodology, 2025, Vol. 21(1), 1–26, https://doi.org/10.5964/meth.14931

Received: 2024-06-25 • **Accepted:** 2025-02-04 • **Published (VoR):** 2025-03-31

Handling Editor: Eduardo Estrada, Autonomous University of Madrid, Madrid, Spain

Corresponding Author: Andres F. Perez Alonso, Department of Methodology and Statistics, Tilburg University, PO Box 90153 5000 LE, Tilburg, the Netherlands. E-mail: A.F.PerezAlonso@tilburguniversity.edu

Supplementary Materials: Code, Materials [see Index of Supplementary Materials]

Abstract

Behavioral scientists often use Multigroup Structural Equation Modeling (MG-SEM) to compare groups in terms of their latent variables (LVs) relations — also called 'structural relations'. Since LVs are measured indirectly, measurement invariance must be evaluated before comparing structural relations. To efficiently compare many groups, the recently proposed Mixture MG-SEM (MMG-SEM) clusters groups based on their structural relations while accounting for measurement (non-)invariance. MMG-SEM requires the user to select the optimal number of clusters for the data at hand. Various approaches address this problem, but no definitive answer exists on which is best. This paper aims to find the best-performing model selection approach for MMG-SEM through a simulation study by comparing five information criteria and the convex hull procedure and including empirically realistic conditions affecting the clusters' separability. No universally best measure was found, but based on our results, we recommend using the convex hull combined with another measure (e.g., AIC) when selecting the number of clusters.

Keywords

model selection, mixture modeling, structural relations, structural equation modeling

Comparing relations between unobservable or 'latent' variables (e.g., attitudes, emotions) across many groups is common in behavioral sciences. For instance, Mayerl and Best (2019) studied how environmental attitudes related to environmental behavior in 30 countries. Structural Equation Modeling (SEM; Bollen, 1989) allows estimating regression coefficients for relations among latent variables (LV) based on the covariances of their observed indicators, such as questionnaire items. In SEM, the regression coefficients are also called 'structural relations' and LVs are called 'factors'.

Multigroup SEM (MG-SEM) and Multilevel SEM (ML-SEM), are commonly used to compare structural relations across groups (e.g., Mayerl & Best, 2019). To pinpoint differences and similarities, these approaches require pairwise comparisons of the group-specific values for the structural relations, which is a complex and daunting task when many groups are involved. For instance, for 30 groups, this would entail 435 pairwise comparisons per parameter. In ML-SEM, the group-specific parameter values are derived from random effects (Hox et al., 2017).

An intuitive solution is to find subsets of groups that share the same relations between factors using mixture modeling (McLachlan et al., 2019). However, before identifying such 'latent classes' or 'clusters', we must remember that the factors are indirectly measured via questionnaire items. Before comparing structural relations between groups, we must ensure that the measurement of the factors is the same across groups or, in other words, that 'measurement invariance' (MI; Meredith, 1993) holds. This implies evaluating whether the measurement model (MM) — indicating which items measure which factors and to what extent — is invariant across groups. The MI assumption can be evaluated at several levels, focusing on different MM parameters (Vandenberg & Lance, 2000). In case of many groups, invariance often does not hold for all the MM parameters. To compare structural relations across groups, the equality of the so-called 'factor loadings' (i.e., the item-factor relations) must hold, which is called metric invariance[1]. Other higher-level MM differences are inconsequential for comparing structural relations if they are included in the model (Chen, 2008; Guenole & Brown, 2014).

When looking for clusters of groups with equal structural relations, capturing the MM differences — or 'measurement non-invariances' — with group-specific parameters is important. Many existing mixture SEM methods (e.g., Kim et al., 2016; Vermunt & Magidson, 2005) force all parameters to be equal within a cluster. This implies that MM parameters can either be specified as invariant across all groups (i.e., ignoring measurement non-invariances) or as cluster-specific (i.e., enforcing MI within each cluster but not across clusters). Such mixture SEM methods capture clusters of groups with the same structural relations as well as the same MM parameters, and fail to disentangle the

1) When full metric invariance does not hold (i.e., all loadings equal), partial metric invariance (Byrne et al., 1989) can be pursued, where some of the loadings are allowed to be different across groups.

differences of interest (i.e., in the structural relations) from those not of interest (i.e., in the MM).

To effectively capture clusters of groups with equivalent structural relations while simultaneously accounting for measurement non-invariances, Perez Alonso et al. (2024) proposed Mixture Multigroup SEM (MMG-SEM), which combines MG-SEM with mixture clustering. Specifically, it combines cluster-specific structural relations with measurement parameters that are partially group-specific, so that the clustering of the groups focuses only on the structural relations, which are of interest to the research question. Note that, essentially, MMG-SEM comprises two different types of LVs: (1) the continuous LVs measured by items at the individual-level, and (2) a categorical LV for the clusters at the group-level.

By gathering groups with equal structural relations in a cluster, MMG-SEM reduces the number of pairwise comparisons needed to pinpoint which relations differ among groups. However, it also introduces a problem inherent to mixture models; that is, for each data set, the appropriate number of clusters must be determined. In empirical research, the 'true' number of clusters is unknown, and the selection of the number of clusters is an important challenge. When too few clusters are selected, one fails to detect potentially interesting differences in the structural relations and, when too many clusters are retained, one ends up with an overly complex model. Perez Alonso et al. (2024) showed that MMG-SEM performs well when the correct number of clusters is specified, but did not address the model selection problem. In their empirical application, they applied a model selection approach recommended for related mixture methods (e.g., De Roover et al., 2022; Lukočiene et al., 2010; Lukočiene & Vermunt, 2009), but they did not evaluate how commonly used model selection approaches perform for MMG-SEM in different conditions or which approach is best for MMG-SEM.

Several approaches to address the model selection problem are available (see Akogul & Erisoglu, 2016). For instance, the Bayesian Information Criterion (BIC; Schwarz, 1978) and Akaike Information Criterion (AIC; Akaike, 1974) integrate model fit and a penalty based on model complexity. Hence, a model that minimizes the criteria is assumed to have a good balance between model fit and parsimony. Another way of finding this balance is using the Convex Hull method (Ceulemans & Kiers, 2006), which is a generalized scree test. Alternatively, the Integrated Completed Likelihood (ICL; Biernacki et al., 2000) also considers the cluster separation; that is, it penalizes models that offer poorly-defined clusters (i.e., clusters that are too similar). This aligns with the fact that substantive researchers likely regard minor differences in structural relations to be trivial.

Numerous simulation studies have compared different model selection methods (e.g., Akogul & Erisoglu, 2016; De Roover et al., 2022; Lukočiene et al., 2010; Lukočiene & Vermunt, 2009; Nylund et al., 2007), showing different results depending on the conditions and mixture models evaluated. For instance, for mixture models combined with factor analysis, Bulteel et al. (2013) and De Roover (2021) found that BIC and Convex

Hull outperformed AIC. In the context of latent class analysis, Lukočiene and Vermunt (2009) found that AIC_3 (i.e., a modified AIC with a larger penalty) performed better than other model selection methods.

These contradictory results emphasize the importance of evaluating and comparing model selection approaches for MMG-SEM specifically, which is the aim of this paper. By means of a simulation study, we will compare different approaches in conditions that mimic the ones found in social sciences. For instance, in empirical data, it is likely that certain groups have very similar — but not identical — regression parameters. Gathering these groups in the same cluster may still be desirable, for the sake of parsimony, and because researchers are often not interested in such trivial differences. Therefore, the simulated conditions will include different levels of small differences in structural relations *within* a cluster, to evaluate how this affects the model selection.

The remainder of this paper is organized as follows: MMG-SEM and relevant model selection methods are described in the Method section. A Simulation Study then evaluates the performance of the model selection methods in the context of MMG-SEM. The paper concludes with a Discussion section highlighting the most relevant results and limitations of the study.

Method

Mixture Multigroup Structural Equation Modeling

Mixture Multigroup Structural Equation Modeling (MMG-SEM; Perez Alonso et al., 2024) combines mixture modeling with MG-SEM. In general, the mixture multigroup approach (De Roover, 2021; De Roover et al., 2022), aims to find a clustering that focuses on specific parameters of interest. In MMG-SEM, the clustering focuses on the structural relations, while MM differences are accounted for by group-specific parameters, so they do not affect the clustering.

For its estimation, Perez Alonso et al. (2024) used the 'Structural-After-Measurement' (SAM; Rosseel & Loh, 2022) approach, which estimates a SEM model in two steps. In the first step, the MM is estimated, whereas the structural model (SM; including the structural relations) is estimated in the second step. The estimation of MMG-SEM is briefly described below (for more details, see Perez Alonso et al., 2024).

Step 1: Measurement Model

The MM defines how the LVs are measured; that is, which items measure which factor and to what extent. When studying multiple groups (e.g., countries), the MM is often estimated using Multigroup Confirmatory Factor Analysis (MG-CFA). If we consider individuals $n_g = 1, ..., N_g$ within groups $g = 1, ..., G$, items $j = 1, ..., J$, and factors $q = 1, ..., Q$, MG-CFA defines the vector of observed scores $\mathbf{x}_{n_g}$ of individual n_g as follows

$$\mathrm{x}_{n_g} = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_{n_g} + \boldsymbol{\epsilon}_{n_g}, \tag{1}$$

where $\boldsymbol{\tau}_g$ is a $\mathcal{J}$-dimensional vector of group-specific intercepts, $\boldsymbol{\Lambda}_g$ denotes a $\mathcal{J} \times Q$ matrix of group-specific factor loadings, $\boldsymbol{\eta}_{n_g}$ is a Q-dimensional random vector of factor scores, and $\boldsymbol{\epsilon}_{n_g}$ is a $\mathcal{J}$-dimensional random vector of residuals. Note that MG-CFA imposes a pattern of zero and non-zero loadings on $\boldsymbol{\Lambda}_g$, and that, in this paper, we center the observed variables per group to remove the mean structure, which is equivalent to estimating the group-specific $\boldsymbol{\tau}_g$, but computationally more efficient. We assume that: (1) $\boldsymbol{\eta}_{n_g}$ is distributed according to a multivariate normal distribution $MVN(\boldsymbol{\alpha}_g, \boldsymbol{\Phi}_g)$, where $\boldsymbol{\alpha}_g$ and $\boldsymbol{\Phi}_g$ are the mean vector and covariance matrix of the factors, respectively, and (2) $\boldsymbol{\epsilon}_{n_g}$ is distributed according to $MVN(0, \boldsymbol{\Theta}_g)$, where $\boldsymbol{\Theta}_g$ is the covariance matrix of the residuals in Group g, which is usually assumed to be diagonal. If we assume that $\mathrm{Cov}(\boldsymbol{\eta}_{n_g}, \boldsymbol{\epsilon}_{n_g}) = \mathbf{0}$, the model-implied covariance matrix of Group g is given by

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Theta}_g\,. \tag{2}$$

In the context of MMG-SEM, one must evaluate measurement invariance (MI) before clustering groups on structural relations. As mentioned in the Introduction, MI can be evaluated at different levels by focusing on different parameters. First, one must evaluate configural invariance by testing if the model in Equation 1 holds across groups. If the model's fit is satisfactory (Chen, 2007), one can assume that the same items measure the same factors across groups. Second, one must test for metric invariance by constraining the non-zero loadings in $\boldsymbol{\Lambda}_g$ to be the same across groups. If imposing $\boldsymbol{\Lambda}_g = \boldsymbol{\Lambda}$ does not significantly worsen the model fit (Rutkowski & Svetina, 2014), full metric invariance holds. Metric invariance must hold, at least partially, for valid comparisons of the structural relations[2]. Therefore, the remaining MM parameters (i.e., $\boldsymbol{\tau}_g$, $\boldsymbol{\Theta}_g$, and potentially some loadings) are allowed to be group-specific in MMG-SEM. If full metric invariance holds, the model-implied covariance matrix of Group g in MMG-SEM's first step is

$$\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda} \boldsymbol{\Phi}_g \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_g\,. \tag{3}$$

The MM is fitted by minimizing the difference between the model-implied covariance matrix $\boldsymbol{\Sigma}_g$ and the observed covariance matrix $\mathbf{S}$. The group-specific factor covariance matrices $\boldsymbol{\Phi}_g$ from Equation 3 are the input for MMG-SEM's second step. To avoid confusion in the notation of the remaining text, the covariance matrices $\boldsymbol{\Phi}_g$ from Step 1 will have a superscript $s1$ (i.e., $\boldsymbol{\Phi}_g^{s1}$).

2) For more information about the remaining MI levels, please see Vandenberg and Lance (2000).

Step 2: Structural Model

In the second step, MMG-SEM estimates the SM and performs the mixture clustering based on the structural relations. Note that MMG-SEM operates at the group-level; that is, it finds clusters of groups instead of clusters of observations. The SM, which defines how the LVs are related, is conditional on the membership of Group g to Cluster k, denoted as z_{gk}, which takes on a value of 1 or 0. Note that the true cluster memberships z_{gk} are unknown and that the estimated $\hat{z}_{gk}$ is a probability ranging from 0 to 1. Formally, the model-implied factor covariance matrix $\mathbf{\Phi}_{gk}$ is defined as:

$$[\mathbf{\Phi}_{gk} | z_{gk} = 1] = (\mathrm{I} - \mathrm{B}_k)^{-1} \mathbf{\Psi}_{gk} (\mathrm{I} - \mathrm{B}_k)^{-1'}, \tag{4}$$

where $\mathbf{B}_k$ is a non-symmetric $Q \times Q$ matrix containing the unstandardized cluster-specific regression coefficients between LVs, and $\mathbf{\Psi}_{gk}$ is the residual factor covariance matrix. The group-and-cluster-specific nature of $\mathbf{\Psi}_{gk}$ ensures that the clustering is driven only by the regression coefficients $\mathbf{B}_k$, and not (also) by the residual factor covariances[3]. The SM for each group-cluster combination gk is fitted by minimizing the differences between the model-implied factor covariance matrices $\mathbf{\Phi}_{gk}$ in Step 2 and the group-specific covariance matrices $\mathbf{\Phi}_g^{s1}$ from Step 1.

For the mixture clustering, MMG-SEM assumes that the vector of factor scores $\boldsymbol{\eta}_{n_g}$ is sampled from a mixture of K multivariate normal distributions and that all factor scores of Group g — gathered in a matrix $\mathbf{H}_g$ of factor scores — are sampled from the same distribution. Specifically, the formal definition for Group g is the following

$$f(\mathbf{H}_g; \vartheta) = \sum_{k=1}^{K} \pi_k f_{gk}(\mathrm{H}_g; \vartheta_{gk}) = \sum_{k=1}^{K} \pi_k \prod_{n_g=1}^{N_g} MVN(\boldsymbol{\eta}_{n_g}; \boldsymbol{\alpha}_g, \mathbf{\Phi}_{gk}), \tag{5}$$

where f is the population density function, ϑ is the set of population parameters, π_k is the prior probability of a Group g belonging to Cluster k (where $\sum_{k=1}^{K} \pi_k = 1$), f_{gk} is the density function of the Group g in the kth cluster, and ϑ_{gk} is its corresponding set of parameters. Specifically, f_{gk} is a multivariate normal distribution where $\mathbf{\Phi}_{gk}$ and $\boldsymbol{\alpha}_g$ are the factors' covariance matrix and mean vector, respectively. The covariance matrix is decomposed as indicated in Equation 4, and the factor means $\boldsymbol{\alpha}_g$ are equal to zero due to the centering.

3) To this aim, the residual (co)variances of the endogenous factors are group-and-cluster-specific, whereas the (co)variances of the exogenous factors are group-specific. For more details, please see the original paper by Perez Alonso et al. (2024).

Model Estimation

The unknown parameters ϑ of Step 2 (Equation 5) are estimated by means of maximum likelihood estimation using an EM algorithm (for details, see Perez Alonso et al., 2024). Specifically, the following log-likelihood function is maximized:

$$\log L_{\boldsymbol{\eta}} = \sum_{g=1}^{G} log\left(\sum_{k=1}^{K} \pi_k \left(\frac{1}{(2\pi)^{Q/2} |\boldsymbol{\Phi}_{gk}|^{1/2}} exp\left(-\frac{1}{2} tr(\boldsymbol{\Phi}_g^{s1} \boldsymbol{\Phi}_{gk}^{-1}) \right) \right)^{N_g} \right), \tag{6}$$

where $\boldsymbol{\Phi}_g^{s1}$ is Step 1's factor covariance (Equation 3), and $\boldsymbol{\Phi}_{gk}$ is Step 2's factor covariance (Equation 4).

The log-likelihood function in Equation 6 considers only the parameters in the SM. The log-likelihood function for the full MMG-SEM model (i.e., combining Step 1 and Step 2) is defined as

$$\log L = \sum_{g=1}^{G} log\left(\sum_{k=1}^{K} \pi_k \prod_{n_g=1}^{N_g} \frac{1}{(2\pi)^{J/2} |\Sigma_{gk}|^{1/2}} exp\left(-\frac{1}{2} (x_{n_g} - \boldsymbol{\mu}_g)' \Sigma_{gk}^{-1} (x_{n_g} - \boldsymbol{\mu}_g) \right) \right), \tag{7}$$

which can also be expressed in terms of covariance matrices to resemble Equation 6 as follows

$$\log L = \sum_{g=1}^{G} log\left(\sum_{k=1}^{K} \pi_k \left(\frac{1}{(2\pi)^{Q/2} |\Sigma_{gk}|^{1/2}} exp\left(-\frac{1}{2} tr(S_g \Sigma_{gk}^{-1}) \right) \right)^{N_g} \right), \tag{8}$$

where $\boldsymbol{\mu}_g$ and Σ_{gk} are the mean vector and model-implied covariance matrix of the observed items, respectively. The $\boldsymbol{\mu}_g$ is zero due to the centering, and the Σ_{gk} can be reconstructed by inserting Equation 4 into Equation 3 as

$$\Sigma_{gk} = \boldsymbol{\Lambda}(I - B_k)^{-1} \boldsymbol{\Psi}_{gk} (I - B_k)^{-1'} \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_g . \tag{9}$$

Finally, step-wise estimation as the one described above brings many advantages, such as an intuitive approach (i.e., deal with the measurement model first and then study the structural parameters of interest), robustness against local model misspecifications (Rosseel & Loh, 2022), and a simplified estimation for complex models (Perez Alonso et al., 2024). However, as explained by Bakk et al. (2014), step-wise estimation also introduces uncertainty in the second step (i.e., additional variance in the estimates), which can lead to biased standard errors if not accounted for. A procedure to correct for biased standard errors is described in detail in Bakk et al. (2014) and Rosseel and Loh (2022). The same procedure is currently implemented in MMG-SEM when computing standard errors (see Perez Alonso, 2025 for details).

Model Selection

A common challenge for clustering methods is selecting an appropriate number of clusters. Researchers have tried to solve this problem based on different approaches, such as balancing model fit and complexity (Akaike, 1974; Akogul & Erisoglu, 2016; Schwarz, 1978), considering relative fit improvement (Ceulemans & Kiers, 2006), cluster separation (Biernacki et al., 2000), and/or substantive interpretation (van den Bergh et al., 2017). Given the popularity of clustering, new methods for model selection keep emerging. However, we focus on commonly used approaches for model selection in the context of mixture SEM methods, which are detailed below.

Akaike Information Criterion

The Akaike Information Criterion (AIC; Akaike, 1974) combines the model fit (i.e., the log-likelihood) with a penalty for model complexity (i.e., number of parameters). It is defined as:

$$\text{AIC} = -2\,\text{logL} + 2P \tag{10}$$

where P is the number of parameters. For MMG-SEM, P is the sum of the number of mixing proportions (minus one restriction), the number of cluster-specific regressions coefficients, the number of group-specific exogenous factor covariances, the number of group-and-cluster-specific endogenous factor covariances[4], the number of loadings (minus Q fixed loadings due to factor scaling and accounting for (non-)invariant loadings), and the number of group-specific unique variances.

A number of modifications of the AIC have been presented. In this paper, we consider only one such modification: the AIC_3 (Bozdogan, 1994), which was developed specifically for determining the number of clusters in mixture models. It is defined as:

$$\text{AIC}_3 = -2\text{logL} + 3P\,. \tag{11}$$

Bayesian Information Criterion

The Bayesian Information Criterion (BIC; Schwarz, 1978) balances model fit and model complexity as follows:

$$\text{BIC} = -2\,\text{logL} + P\text{log}(SS), \tag{12}$$

where the penalty of the model complexity is now weighted by the logarithm of the sample size SS. Usually, the total number of observations N is used as the SS (BIC_N), but

4) We only count one set of endogenous covariances for each group, given we assume each group belongs to only one cluster. The endogenous covariances (from the clusters the groups do not belong to) are nuisance parameters.

it has been suggested to use the number of Groups G instead of N (BIC_G) when selecting the number of group-level clusters (Lukočienė et al., 2010; Lukočienė & Vermunt, 2009). De Roover (2021) and De Roover et al. (2022) found a superior performance of BIC_G in the context of a related mixture multigroup approach.

Convex Hull

The convex hull procedure (CHull; Ceulemans & Kiers, 2006) has been shown to be a valid alternative to BIC and AIC in the context of mixtures of factor analyzers (Bulteel et al., 2013). CHull is a generalized scree test that balances model fit and model complexity by plotting the *logL* of the different models in function of their number of parameters *P*. Then, for each model on convex hull of the scree plot, a scree ratio is computed and the solution with the maximal scree ratio is selected. Specifically, the scree ratio sr_d for model d is defined as:

$$sr_d = \frac{\text{logL}_d - \text{logL}_{d-1}}{P_d - P_{d-1}} \Big/ \frac{\text{logL}_{d+1} - \text{logL}_d}{P_{d+1} - P_d}, \tag{13}$$

where $d - 1$ refers to the previous (less complex) model on the hull and $d + 1$ refers to the next (more complex) model on the hull. It is worth noting that a scree ratio cannot be computed for the least complex model, so it will always select a model with at least two clusters. However, if no clear elbow in the scree plot, one may still conclude that an underlying cluster is unlikely.

Integrated Completed Likelihood

The Integrated Completed Likelihood (ICL; Biernacki et al., 2000) is a model selection criterion developed for mixture clustering as an alternative to the BIC. The BIC does not consider an essential aspect of the mixture models; that is, the estimated cluster memberships $\hat{z}_{gk}$. Therefore, Biernacki et al. (2000) proposed using the *Entropy*, which is a measure of the uncertainty of Group g belonging to Cluster k. Remember that $\hat{z}_{gk}$ is a probability ranging from 0 to 1. Formally, for MMG-SEM, the *Entropy* can be defined as:

$$Entropy = \sum_{g=1}^{G} \sum_{k=1}^{K} (-\hat{z}_{gk})(\log \hat{z}_{gk}) \tag{14}$$

The complete derivation of the ICL can be found in Biernacki et al. (2000), but its approximation, based on the BIC, is rather simple. Formally, the ICL is approximated as[5]:

5) Biernacki et al. (2000) defined the ICL slightly differently as BIC − *Entropy*, but they defined BIC = logL − $\frac{P}{2}$log(*SS*) instead of BIC = −2 logL + *P*log(*SS*). Therefore, both ICL definitions will lead to the same results in terms of model selection.

$$\text{ICL} = \text{BIC} + 2Entropy \tag{15}$$

For brevity, we focus on the ICL based on BIC_G in the Simulation Study, since this has been shown to perform better than BIC_N in the context of group-level clustering.

Simulation Study

Design

The aim of the simulation study was to compare the performance of six model selection measures in selecting the number of clusters for MMG-SEM: AIC, AIC_3, BIC_G, BIC_N, CHull, and ICL. To this end, we used a Monte-Carlo simulation with six manipulated factors that are expected to affect the model selection performance. The factors and their corresponding levels are described below:

1. Size of regression parameters β: 0.3, 0.4;
2. Number of groups G: 24, 48;
3. Within-group sample size N_g: 50, 100, 200;
4. Number of clusters K: 2, 4;
5. Cluster size: balanced, unbalanced;
6. Within-cluster differences σ_β: no difference (0), small (0.05), large (0.1).

The size of the regression parameters (i.e., 0.3 and 0.4) was inspired by previous simulation studies in SEM (e.g., Guenole & Brown, 2014; Perez Alonso et al., 2024) and cross-national empirical studies with many groups (e.g., Bastian et al., 2014; Kuppens et al., 2008). Similarly, the number of groups was 24 and 48, which correspond to a range of groups that is generally found in empirical large-scale international surveys (Rutkowski & Svetina, 2014). The number of clusters (i.e., 2 and 4)[6] and cluster size were defined considering previous simulation studies on model selection (Biernacki et al., 2000; Bulteel et al., 2013; De Roover, 2021; De Roover et al., 2022; Lukočienė et al., 2010; Perez Alonso et al., 2024; Steinley & Brusco, 2011) as we want to relate our results to theirs.

In total, the design included 2 (size of regression parameters) × 2 (number of groups) × 3 (within-group sample size) × 2 (number of clusters) × 2 (cluster size) × 3 (within-cluster differences) = 144 data generation conditions. For all conditions, 100 different data sets were generated, for a total of 14400 data sets. To evaluate the model selection measures, each data set was analyzed six times with MMG-SEM from one to six clusters, for a total of 14400 × 6 = 86400 analyses. Note that we added non-invariances to the loadings (see the Data Generation section for more information) and that such non-invariances

6) A smaller simulation where $K = 1$ was also performed but not included in the main text due to space constraints. More details can be found in Perez Alonso et al. (2025)

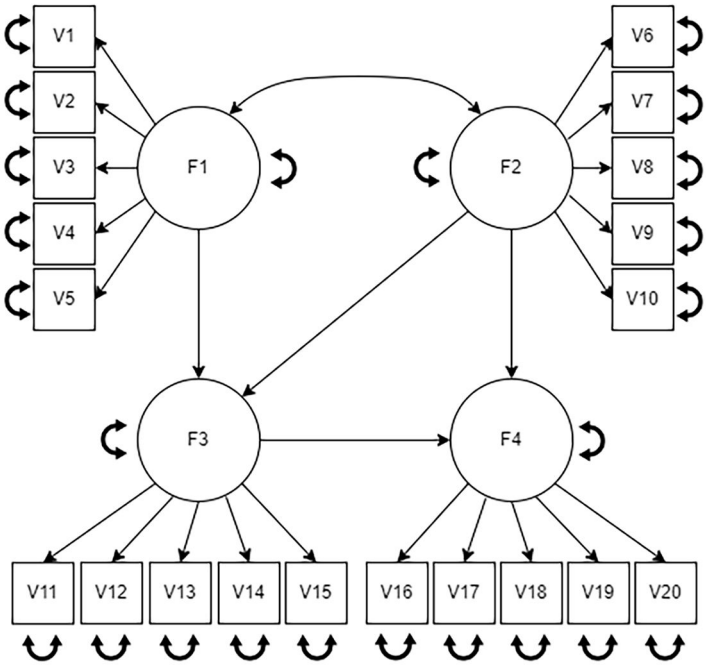
are correctly modeled in MMG-SEM. All the data generation and analyses were done using R Version 4.3.3 (R Core Team, 2024), and the code is openly shared at Perez Alonso (2024). The data generation procedure is described below.

Data Generation

Each data set was generated according to the SEM model in Figure 1, with four LVs, each one measured by five indicators, for a total of 20 observed variables. The structural relations, which are the parameters of interest for the clustering, are represented by four regression parameters. F1 and F2 served as exogenous variables, while F3 and F4 were endogenous variables. Note that F3 acts as a 'mediator' and is, thus, an exogenous and endogenous variable at the same time.

Figure 1

The Model Used for the Data Generation



Note. F1 and F2 are exogenous variables, F3 is dependent and independent at the same time ('mediator'), and F4 is a dependent only variable.

The sample size per Group N_g, the number of Groups G, and the number of clusters K were defined according to the manipulated factors 'within-group sample size', 'number

of groups', and 'number of clusters', respectively. The number of groups per cluster was manipulated according to the 'cluster size'. In the balanced condition, the groups were equally divided per cluster. For instance, for $G = 48$ and $K = 4$, each cluster contained 12 groups. In contrast, in the unbalanced condition, there was one larger cluster with 75% of the groups, and the remaining clusters were equally sized. For example, for $G = 48$ and $K = 4$, the first cluster would contain 36 groups and the remaining three clusters would contain four groups each.

The 20 observed variables were generated from a $MVN(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{gk})$, where the mean vector $\boldsymbol{\mu}_g$ was a vector of zeros and the covariance matrices $\boldsymbol{\Sigma}_{gk}$ were generated according to Equation 9. Thus, to generate the data, the parameters in Equation 9 (i.e., $\boldsymbol{\Lambda}, \boldsymbol{\Theta}_g, \mathbf{B}_k$, and $\boldsymbol{\Psi}_{gk}$) must be defined. The $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}_g$ matrices were generated aiming to obtain a total variance per item around 1 and a reliability (R^2) of 0.6 for each observed indicator. To do this, the non-zero values in $\boldsymbol{\Lambda}$ were set to $\sqrt{0.6}$ while the residual variances in $\boldsymbol{\Theta}_g$ were drawn from a uniform distribution $U(0.3, 0.5)$. Note that the residual variances in $\boldsymbol{\Theta}_g$ were sampled for each group, allowing group-specific differences as specified in Equation 9.

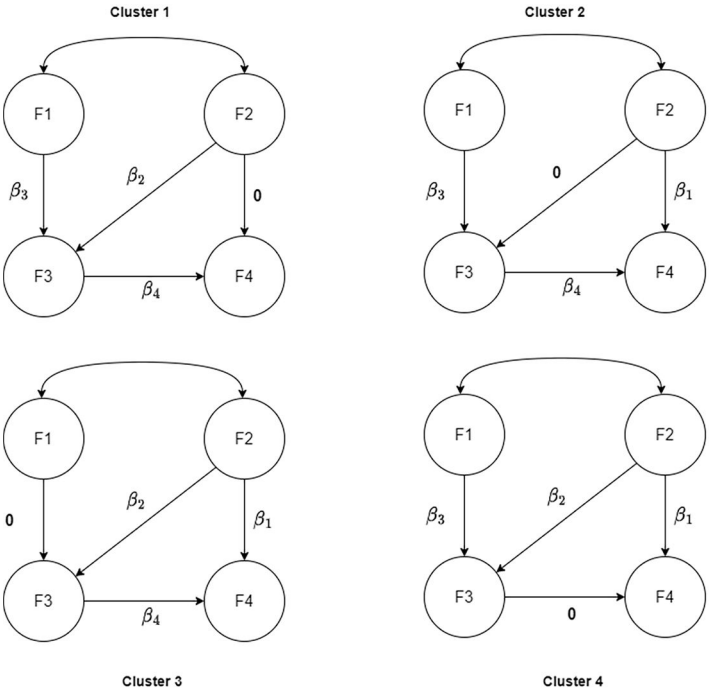
To evaluate the effect of loading non-invariances on the model selection, we also added between-group differences to the $\boldsymbol{\Lambda}$ matrices. In particular, 50% of the groups presented non-invariances. For each non-invariant group, we applied the non-invariance to the second and third loading of each factor (the first loading is fixed to 1). The non-invariances were randomly sampled from a uniform distribution around 0.4, i.e., $U(0.3, 0.5)$, and it was randomly decided whether the non-invariance was added or subtracted to the original loading (i.e., $\sqrt{0.6}$). As a result, each non-invariant loading was different for each non-invariant group.

The setup of the regression parameters in $\mathbf{B}_k$ can be seen in Figure 2. The manipulated factor 'size of the regression parameters' (β) indicated the size of the coefficients β_1, β_2, β_3, and β_4. The difference between the clusters was created by setting one of those coefficients to zero in each cluster. Thus, the size of the regression parameters also defined the size of the difference between the clusters. Note that, when $K = 2$, Models 3 and 4 in Figure 2 were not applicable. The parameters in $\mathbf{B}_k$ were also affected by the manipulated factor 'within-cluster differences' (σ_β). To simulate empirically realistic conditions, we added small differences in the coefficients β to each Group g within a Cluster k. To do this, within each cluster, the regression parameter of each group was drawn from a normal distribution $N(\beta, \sigma_\beta)$, where the β and the σ_β acted as the mean and the standard deviation of the distribution, respectively. The values of σ_β implied no within-cluster differences ($\sigma_\beta = 0$), small differences ($\sigma_\beta = 0.05$), or large differences ($\sigma_\beta = 0.1$). Note that we expect 99% of the sampled values to be within three standard deviations of the mean when drawing from a normal distribution. For instance, if we consider $\beta = 0.3$ and $\sigma_\beta = 0.05$ and we focus on β_1 (see Figure 2), the values of β_1 in Cluster 1 will be sampled from $N(0, 0.05)$, whereas they will be sampled from $N(0.3, 0.05)$ in Cluster 2. We expect 99% of the values to lie between -0.15 and 0.15 in Cluster 1 and

between 0.15 and 0.45 in Cluster 2. Thus, the small differences level ($\sigma_\beta = 0.05$) led to almost no overlap between clusters, whereas the large level ($\sigma_\beta = 0.1$) ensured overlap between the clusters.

Figure 2

Zero and Non-Zero Regression Parameters Between the LVs Depending On the Cluster



Finally, the elements in $\mathbf{\Psi}_{gk}$ were defined by sampling the variance of the exogenous variables F1 and F2 from a uniform distribution $U(0.75, 1.25)$, and their covariance from $U(-0.3, 0.3)$ for all groups. Similarly, the total variance of the endogenous factors F3 and F4 was sampled from $U(0.75, 1.25)$ for each group, and their residual variance depended on the cluster-specific regression parameters. For example, if the total variance of F3 for Group g was Var_{tot}, the residual variance Var_{res} for Group g and Cluster k was $Var_{res} = Var_{tot} - (\beta_2^2 \mathrm{Var}(F1) + \beta_3^2 \mathrm{Var}(F2) + 2\,\beta_2\,\beta_3\,\mathrm{Cov}(F1, F2))$.

R^2 Entropy

To determine the cluster separation in the simulated datasets, we applied the R^2 Entropy, an *Entropy*-based measure. Specifically, the R^2 Entropy indicates how well the observed

responses predict the cluster memberships $\hat{z}_{gk}$ (for details on its calculation, see Vermunt & Magidson, 2016). It takes on a value of 1 when the clusters are perfectly separated (i.e., no classification uncertainty) and a value of 0 when there is no separation at all.

To get an overview of how the cluster separation was affected by the simulation conditions, we evaluated the R^2 Entropy at (an approximated) population level. For this, we generated data for each simulation condition with a large sample size. Given that MMG-SEM's clustering is at the group-level, the relevant sample size for the clustering and R^2 Entropy is the number of groups G, which was set to 192 groups. Subsequently, we computed the cluster memberships $\hat{z}_{gk}$ based on the true parameters values for $\mathbf{\Lambda}, \mathbf{\Theta}_g, \mathbf{B}_k$, and $\mathbf{\Psi}_{gk}$ (i.e., they were used as starting values in an MMG-SEM analysis where no parameter updates were performed). Per condition, 300 replications were generated.

The R^2 Entropy ranged from 0.76 to 1 with an average of 0.93 across all data sets, which indicated well-separated clusters overall. The R^2 Entropy was mostly influenced by the within-cluster differences σ_β, the regression coefficients β, and the within-group sample size N_g. Their main effects can be seen in Table 1, and the interaction between σ_β and β is shown in Figure 3. Lower R^2 Entropy values were found in difficult conditions involving large within-cluster differences ($\sigma_\beta = 0.1$), lower regression coefficients (β = 0.3) and/or low within-group sample size ($N_g = 50$). It is expected that the model selection measures will struggle to choose the correct number of clusters in conditions where the clusters are not well separated.

Table 1

Average R^2 Entropy per Level of the Most Influential Factors

N_g			σ_β			β	
50	**100**	**200**	**0**	**0.05**	**0.1**	**0.3**	**0.4**
0.923	0.982	0.993	0.983	0.976	0.938	0.942	0.990

Figure 3

Approximated Population R^2 Entropy in Function of the Within-Cluster Differences σ_β and the Size of the Regression Parameters β

Results

Before discussing the findings, it is worth noting that the evaluated model selection methods all use the log-likelihood as the measure of model fit. For MMG-SEM, we could use the log-likelihood based on the factors (Equation 6) or the observed data log-likelihood (Equation 7). We evaluated the measures' performance using both log-likelihoods in our simulation and found only minor differences in the results. Thus, for brevity, we present only the results using the log-likelihood in Equation 7, since this considers how the full model (i.e., measurement + structural model) fits the data. The results using Equation 6 can be found in Perez Alonso et al. (2025).

Model Selection

For each model selection measure, we assessed how often it correctly selected the true number of clusters. To gain a deeper understanding of each measure, we also inspected how often it over- and under-selected the number of clusters. The main effects of the manipulated factors of the simulation can be seen in Table 2. In total, the best-performing method was the CHull, followed by the AIC, AIC_3, BIC_G, ICL, and BIC_N, with a proportion of correctly selected models of 0.77, 0.66, 0.64, 0.63, 0.61, and 0.54, respectively. Note that the similarity between the results of AIC_3 and BIC_G can be partially explained by the similarity of their penalties (see Equations 11 and 12); that is, AIC_3's penalty is 3, whereas BIC_G's penalty is 3.18 and 3.87 when G is 24 and 48, respectively.

Table 2

Proportion of Under-, Over-, and Correct Selection of the Number of Clusters for All Model Selection Measures Per Level of Each Manipulated Factor

		K		N_g			G		β		Cluster size		σ_β			
Measure	**Result**	**2**	**4**	**50**	**100**	**200**	**24**	**48**	**0.3**	**0.4**	**Bal**	**Unb**	**0**	**0.05**	**0.1**	**Total**
AIC	Under	0.02	0.34	0.44	0.10	0.00	0.22	0.14	0.25	0.11	0.12	0.25	0.19	0.19	0.17	0.18
	Correct	0.82	0.50	0.56	0.76	0.66	0.64	0.68	0.61	0.71	0.73	0.59	0.80[a]	0.81	0.37	0.66
	Over	0.16	0.16	0.00	0.13	0.34	0.14	0.18	0.14	0.18	0.16	0.16	0.01	0.01	0.46	0.16
AIC_3	Under	0.05	0.41	0.52	0.17	0.01	0.30	0.17	0.32	0.15	0.17	0.30	0.24	0.24	0.22	0.23
	Correct	0.82	0.46	0.48	0.78	0.67	0.60	0.69	0.57	0.71	0.71	0.57	0.75	0.76	0.42	0.64
	Over	0.12	0.12	0.00	0.05	0.32	0.10	0.14	0.11	0.14	0.12	0.13	0.01	0.00	0.36	0.12
BIC_G	Under	0.07	0.45	0.57	0.20	0.01	0.31	0.21	0.35	0.17	0.19	0.33	0.28	0.27	0.24	0.26
	Correct	0.82	0.44	0.43	0.77	0.69	0.59	0.66	0.55	0.70	0.70	0.56	0.72	0.73	0.43	0.63
	Over	0.12	0.11	0.00	0.04	0.30	0.10	0.13	0.10	0.13	0.11	0.12	0.01	0.00	0.33	0.11
BIC_N	Under	0.17	0.70	0.74	0.42	0.15	0.49	0.38	0.57	0.30	0.36	0.51	0.44	0.44	0.43	0.44
	Correct	0.81	0.28	0.26	0.58	0.79[a]	0.50	0.59	0.42	0.66	0.62	0.47	0.56	0.56	0.51	0.54
	Over	0.03	0.02	0.00	0.00	0.06	0.01	0.03	0.01	0.03	0.02	0.02	0.00	0.00	0.06	0.02
Chull	Under	0.00	0.20	0.14	0.08	0.08	0.10	0.10	0.12	0.07	0.06	0.14	0.08	0.07	0.15	0.10
	Correct	0.89[a]	0.66[a]	0.73[a]	0.81[a]	0.78	0.76[a]	0.78[a]	0.74[a]	0.80[a]	0.86[a]	0.68[a]	0.79	0.86[a]	0.67[a]	0.77[a]
	Over	0.11	0.15	0.14	0.11	0.13	0.14	0.12	0.13	0.12	0.08	0.18	0.13	0.07	0.18	0.13
ICL	Under	0.10	0.47	0.63	0.21	0.02	0.33	0.24	0.39	0.18	0.22	0.35	0.30	0.29	0.27	0.28
	Correct	0.79	0.43	0.37	0.77	0.70	0.58	0.65	0.53	0.70	0.68	0.55	0.70	0.71	0.43	0.61
	Over	0.11	0.10	0.00	0.03	0.28	0.09	0.11	0.08	0.12	0.10	0.10	0.00	0.00	0.30	0.10
Total	Under	0.07	0.43	0.51	0.20	0.05	0.29	0.21	0.33	0.16	0.19	0.31	0.25	0.25	0.24	0.25
	Correct	0.82	0.46	0.47	0.74	0.72	0.61	0.67	0.57	0.72	0.72	0.57	0.72	0.74	0.47	0.64
	Over	0.11	0.11	0.02	0.06	0.24	0.10	0.12	0.10	0.12	0.10	0.12	0.03	0.01	0.28	0.11

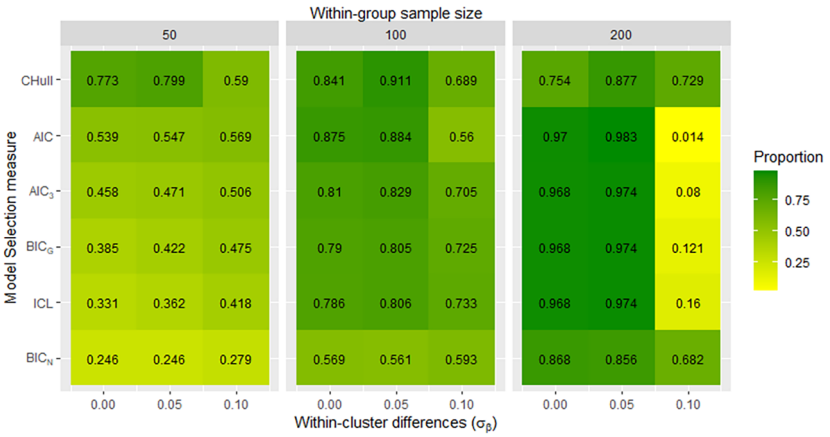
Note. K is the number of clusters, N_g is the within-group sample size, G is the number of groups, β is the size of the regression coefficients, *Bal* is balanced, *Unb* is unbalanced, and σ_β.
[a]Denotes best results.

The within-group sample size N_g was most influential on the model selection performance. Specifically, on average, the model selection measures were correct 47% and 74% of the times when $N_g = 50$ and $N_g = 100$, respectively. Such dramatic improvement did not hold when N_g increased to 200, for which the measures were correct 72% of the time. The improvement from $N_g = 50$ to $N_g = 100$ aligns with common sample size requirements in SEM, since 100 is considered the minimum for consistent estimates (Gorsuch, 1983). The slight decrease in performance when $N_g = 200$ can be explained by the trends of under- and over-selection of the number of clusters. Generally, when choosing the incorrect model, the measures tended to under-select. However, over-selection was more prominent in the case of a large within-group sample size ($N_g = 200$). Such results are unsurprising, considering that larger sample sizes give more power to identify smaller

differences (leading to more clusters). This is more likely in case of larger within-cluster differences ($\sigma_\beta = 0.1$), which can be identified as additional clusters. This can be clearly seen in Figure 4, where the model selection performance dramatically drops when $N_g = 200$ and $\sigma_\beta = 0.1$.

Figure 4

Proportion of Correctly Selected Models in Function of the Within-Group Sample Size N_g and the Within-Cluster Differences σ_β



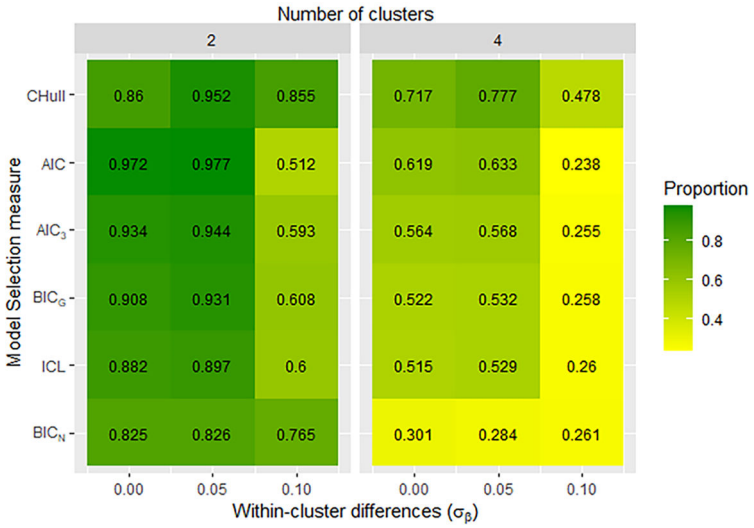
The model selection performance was also greatly affected by the number of clusters K. On average, all model selection measures found it more difficult to identify the correct model when more clusters were underlying the data. From Table 2, the proportion of correctly selected models was better when $K = 2$ (0.82) than when $K = 4$ (0.46). Such results are in line with previous research where an increase in the true number of clusters dramatically decreased the performance of the model selection measures (Bulteel et al., 2013; Lukočienė et al., 2010) even when the cluster separation is high (Steinley & Brusco, 2011). Similarly, the effect of the cluster size on model selection showed a comparable trend. On average, the model selection measures were correct 72% of the time when the cluster size was balanced and 57% when it was unbalanced. The effect of both the number of clusters and cluster size can be explained by lower within-cluster sample sizes in case of more and/or unbalanced clusters, which reduced the power to detect the appropriate model. The model selection was also, to a lesser extent, affected by the regression coefficients β. Specifically, a lower regression coefficient and unbalanced cluster sizes lowered the proportion of correctly selected models.

The interaction between two of the most important factors can be seen in Figure 5. The plot clearly shows that more clusters and large within-cluster differences led to

a dramatic decrease in the performance of all model selection measures. Specifically, the AIC, AIC_3, BIC_G, and ICL presented a substantial decrease of the correctly selected number of clusters. In contrast, the CHull and BIC_N presented a lower decrease in performance when $K = 4$ and/or $\sigma_\beta = 0.1$. BIC_N's performance was generally worse than all other model selection measures, however.

Figure 5

Proportion of Correctly Selected Models in Function of the Number of Clusters K and the Within-Cluster Differences σ_β



From Figures 4 and 5, we also learn that the performance sometimes improved when going from $\sigma_\beta = 0$ to $\sigma_\beta = 0.05$, especially for the CHull. For the CHull, this can be explained by the saturation effect, which happens when adding more clusters results in a negligible increase in the log-likelihood. This can lead to an artificially large scree ratio (because the denominator approaches zero, see Equation 13), whereas, when looking at the scree plot, a virtually horizontal line, rather than an elbow, is visible at this point. In empirical practice, one can remedy this problem by looking for a clear elbow in the scree plot instead of just relying on the scree ratios.

The comparison between CHull and the other measures can be considered unfair, since CHull selects a model with at least two clusters. Thus, when $K = 2$, the other model selection measures may select a one-cluster model when the cluster separation is low, while CHull will select at least two clusters [7]. For a fairer comparison, we also checked the results for the other model selection measures when considering only the models from two to six clusters. In this case, AIC, AIC_3, BIC_G, ICL, and BIC_N selected the right

number of clusters for 66.9%, 66.9%, 66%, 66.4%, and 62.8% of the datasets, respectively. Thus, their performance was closer to (but still lower than) that of the CHull (77%).

Finally, for a more comprehensive understanding of the outcomes, we examined how often the model selection measures correctly identified the number of clusters when considering the two best models (e.g., how often is the correct model among the two models with the lowest AIC values). The correct model was among the two best models 78.5%, 72.7%, 73.1%, 72.9%, 73.1%, and 69.3% of the times for CHull, AIC, AIC_3, BIC_G, ICL, and BIC_N, respectively. The BIC_G, BIC_N, and ICL showed the largest improvements in performance compared to when we focus only on the best model.

Cluster Recovery

The model selection results must be interpreted in light of the cluster recovery (i.e., to what extent MMG-SEM assigns groups to the correct clusters when the correct number of clusters is specified). To this end, we computed the Adjusted Rand Index (ARI; Hubert & Arabie, 1985) for the model with the true number of clusters. For instance, if $K = 2$, we computed the ARI for the model with two clusters (regardless of what the model selection measures chose). The ARI compares two partitions (i.e., modal assignments of the groups to a cluster), taking on a value of 1 for complete agreement and 0 when agreement does not exceed that between two random partitions. Note that it can take on negative values if the agreement is less than what is expected at random. The average ARI across all conditions was good (0.93), but it ranged from -0.10 to 1. Per model selection measure, we also inspected the average ARI across data sets depending on whether the number of clusters was under-, over-, or correctly selected (Table 3). Clearly, selecting the incorrect model (i.e., under- or over-selection) was related to the ARI being lower for the correct model. Specifically, the average ARI was below 0.89 for all measures when the selected model was incorrect, while it was above 0.97 when the correct model was selected.

Table 3

Average ARI for the Model With the True Number of Clusters Depending On Whether the Number of Clusters Was Correctly Selected or Over- or Under-Selected by the Model Selection Measures

Result	AIC	AIC_3	BIC_G	BIC_N	CHull	ICL
Under	0.83 (0.19)	0.85 (0.19)	0.86 (0.19)	0.89 (0.18)	0.77 (0.22)	0.87 (0.18)
Correct	0.98 (0.07)	0.98 (0.07)	0.98 (0.07)	0.98 (0.08)	0.97 (0.09)	0.98 (0.07)
Over	0.86 (0.20)	0.85 (0.20)	0.85 (0.20)	0.89 (0.18)	0.85 (0.20)	0.85 (0.21)

Note. The standard deviation is in parentheses.

7) As can be seen in Table 2, CHull never under-selects the number of clusters when $K = 2$

Section Conclusion

Before drawing conclusions, it is important to note that, strictly speaking, there are as many clusters as there are groups when $\sigma_\beta > 0$ (i.e., in case of within-cluster differences). As MMG-SEM does not capture within-cluster differences, the model selection could suggest extracting more clusters or even capturing each group as a separate cluster, especially when there is enough power to find small differences. However, in this study, we still assumed the true number of clusters to be K, since it is desirable to assign groups with very similar regression parameters to the same cluster (see Introduction). Considering every group as a separate cluster would boil back down to an MG-SEM with group-specific relations and the pesky pairwise comparisons thereof. As this is what we wanted to avoid, we did not include the MG-SEM model in this study.

In an extensive simulation study, we assessed six different model selection measures for MMG-SEM. As expected, lower within-group sample sizes ($N_g = 50$), large within-cluster differences ($\sigma_\beta = 0.1$), more clusters ($K = 4$), and unbalanced cluster sizes decreased the performance of all measures. Overall, the best-performing measure was the CHull and the worst was the BIC_N. AIC, AIC_3, BIC_G, and ICL presented similar performances with minor differences depending on specific conditions.

Considering our results, we suggest using the CHull when performing model selection for MMG-SEM. Since it cannot select the minimum or maximum number of clusters, we suggest also inspecting CHull's scree plot and looking for an elbow to confirm the number of clusters. If no elbow is visible, the most appropriate number of clusters is likely one or the maximum number evaluated. Furthermore, since CHull was not universally best across all data sets[8], we recommend combining its results with at least one of the other measures (e.g., AIC, AIC_3 or BIC_G) to validate a decision. For instance, selecting $K = 1$ when no elbow is found for CHull and AIC suggests a one-cluster model (a small simulation study about model selection performance when $K = 1$ can be found in Perez Alonso et al., 2025). It may also help to consider the best solution for the different measures and compare the solutions in terms of which differences in structural relations are found (and which clustering) and how this relates to prior theories and previous research about the matter.

General Discussion

When using MMG-SEM, the user must specify the appropriate number of clusters for a given data set, as is the case for all clustering techniques. However, the 'true' number of clusters is typically unknown when dealing with real-world data. Therefore, researchers

8) For instance, out of the total 14400 data sets: (1) AIC and CHull were both correct in 8377 cases; (2) AIC was incorrect and CHull correct in 2754 cases; and (3) AIC was correct and CHull incorrect in 1105 cases.

often rely on model selection measures to decide on the number of clusters. Several model selection measures have been evaluated for other clustering methods, but there is no clear-cut answer to which measure is the best one. Different results were found depending on the clustering method, the conditions assessed, and the level at which the clustering is performed (i.e., observation or group level) (Akogul & Erisoglu, 2016; De Roover et al., 2022; Lukočienė et al., 2010; Nylund et al., 2007). Considering the conflicting results and the unique properties of MMG-SEM, such as the combination of group- and cluster-specific parameters, and a clustering focused on regression parameters, prior conclusions on their model selection performance cannot be generalized to MMG-SEM. Therefore, this paper aimed to find the best-performing model selection measure for MMG-SEM through an extensive simulation study. In particular, we compared six model selection measures (i.e., CHull, AIC, AIC_3, BIC_G, ICL, and BIC_N), and included conditions that affect the cluster's separability and mimic empirically realistic conditions. In particular, the small within-cluster differences resembled the small (and trivial) differences between groups that will often be found in empirical research but that are ineffectual to the substantive conclusions on how structural relations differ.

Overall, the best-performing measure was the CHull, followed by the AIC, AIC_3, BIC_G, ICL, and BIC_N. While, in general, this is in line with some previous studies (e.g., Bulteel et al., 2013; De Roover, 2021; De Roover et al., 2022), the clear difference in performance between CHull and the other measures was a remarkable find. This difference could not be explained by the fact that CHull can only select at least two clusters. CHull's advantage may result from its flexibility and lack of assumptions compared to the other measures. For instance, some argue that the true model must be among the candidates for BIC to have consistent results (Vrieze, 2012), which was, strictly speaking, not always the case in the simulation study since the 'true' model was one with group-specific (instead of cluster-specific) regression parameters in case of within-cluster differences.

The vastly inferior performance of BIC_N coincides with previous results for mixture models at the group level that showed that using the number of observations N instead of G as the sample size in the BIC leads to overpenalization and, thus, under-selection (De Roover, 2021; De Roover et al., 2022; Lukočienė et al., 2010; Lukočienė & Vermunt, 2009). More surprising was the superior performance of the AIC over the BIC_G, given that previous simulations have shown a slight advantage of BIC_G in latent class and mixture models (De Roover, 2021; Lukočienė et al., 2010; Lukočienė & Vermunt, 2009). In other studies, AIC slightly outperformed the BIC_G, however (De Roover et al., 2022).

It is worth mentioning that simulation-based research always comes with limitations. Specifically, the results cannot be straightforwardly generalized to conditions that were not assessed in the simulation. For instance, we only included data and models with continuous and normally distributed items, whereas earlier simulations showed different results depending on the type of indicator (Lukočienė et al., 2010). Since data in

social sciences commonly uses ordinal indicators and often presents non-normality (e.g., skewness), it is important to extend the model selection evaluations to MMG-SEM with ordinal indicators and robust estimators for non-normality. Note that the CHull could be computed with measures of model fit other than log-likelihood (e.g., the distance between the model-implied and observed covariance matrices) and avoid the distributional assumptions that come with it. Moreover, in the Simulation Study, we focused on model selection measures that are most commonly used for mixture models in social sciences, overlooking other measures. For instance, the Kullback Information Criterion (Cavanaugh, 1999) is a promising alternative for large sample conditions (Akogul & Erisoglu, 2016); the normalized information criteria (Cohen & Berchenko, 2021) was developed for the presence of missing data; and the Likelihood Increment Percentage per Parameter (Grimm et al., 2021; McArdle et al., 2002) provides a 'effect size' measure of the relative fit improvement in the context of model selection, and is thus an alternative to CHull.

Finally, for MMG-SEM — which compares structural relations between groups using clusters — selecting an appropriate number of clusters is essential for the research questions. Indeed, under- or over-selecting clusters may impair the conclusions on differences and similarities in the relations of interest. When selecting too few clusters, important differences may be overlooked. When selecting too many clusters, one ends up with an overly complex model that implies more comparisons of cluster-specific regression coefficients and, thus, a higher risk of false positives. Therefore, we are happy to conclude that the model selection measures assessed in this paper offer promising solutions to the model selection problem for MMG-SEM, especially when combined (e.g., CHull and AIC).

Funding: This research was funded by a Vidi grant [VI.Vidi.201.133] awarded to Kim De Roover by the Netherlands Organization for Scientific Research (NWO).

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article, the following Supplementary Materials are available:

- Code. (Perez Alonzo, 2024)
- Study materials. (Perez Alonso et al., 2025)

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Akogul, S., & Erisoglu, M. (2016). A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications, 21*(3), Article 34. https://doi.org/10.3390/mca21030034

Bakk, Z., Oberski, D. L., & Vermunt, J. K. (2014). Relating latent class assignments to external variables: Standard errors for correct inference. *Political Analysis, 22*(4), 520–540. https://doi.org/10.1093/pan/mpu003

Bastian, B., Kuppens, P., De Roover, K., & Diener, E. (2014). Is valuing positive emotion associated with life satisfaction? *Emotion, 14*(4), 639–645. https://doi.org/10.1037/a0036466

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(7), 719–725. https://doi.org/10.1109/34.865189

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. In H. Bozdogan, S. L. Sclove, A. K. Gupta, D. Haughton, G. Kitagawa, T. Ozaki & K. Tanabe (Eds.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An informational approach* (pp. 69–113). Springer Netherlands. https://doi.org/10.1007/978-94-011-0800-3_3

Bulteel, K., Wilderjans, T. F., Tuerlinckx, F., & Ceulemans, E. (2013). CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods, 45*(3), 782–791. https://doi.org/10.3758/s13428-012-0293-y

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

Cavanaugh, J. E. (1999). A large-sample model selection criterion based on Kullback's symmetric divergence. *Statistics & Probability Letters, 42*(4), 333–343. https://doi.org/10.1016/S0167-7152(98)00200-4

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology, 59*(1), 133–150. https://doi.org/10.1348/000711005X64817

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Cohen, N., & Berchenko, Y. (2021). Normalized information criteria and model selection in the presence of missing data. *Mathematics, 9*(19), Article 2474. https://doi.org/10.3390/math9192474

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with Mixture Multigroup Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods, 27*(3), 281–306. https://doi.org/10.1037/met0000355

Gorsuch, R. (1983). *Factor Analysis* (2nd ed.). Lawrence Erlbaum.

Grimm, K. J., Houpt, R., & Rodgers, D. (2021). Model fit and comparison in Finite Mixture Models: A review and a novel approach. *Frontiers in Education, 6*, Article 613645. https://doi.org/10.3389/feduc.2021.613645

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology, 5*, Article 980. https://doi.org/10.3389/fpsyg.2014.00980

Hox, J. J., Moerbeek, M., & Van De Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218. https://doi.org/10.1007/BF01908075

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using Multilevel Factor Mixture Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 870–887. https://doi.org/10.1080/10705511.2016.1196108

Kuppens, P., Realo, A., & Diener, E. (2008). The role of positive and negative emotions in life satisfaction judgment across nations. *Journal of Personality and Social Psychology, 95*(1), 66–75. https://doi.org/10.1037/0022-3514.95.1.66

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower- and higher-level classes in Multilevel Latent Class Analysis. *Sociological Methodology, 40*(1), 247–283. https://doi.org/10.1111/j.1467-9531.2010.01231.x

Lukočienė, O., & Vermunt, J. K. (2009). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–249). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-01044-6_22

Mayerl, J., & Best, H. (2019). Attitudes and behavioral intentions to protect the environment: How consistent is the structure of environmental concern in cross-national comparison? *International Journal of Sociology, 49*(1), 27–52. https://doi.org/10.1080/00207659.2018.1560980

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology, 38*(1), 115–142. https://doi.org/10.1037/0012-1649.38.1.115

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application, 6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525–543. https://doi.org/10.1007/BF02294825

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Perez Alonzo, A. F. (2024). *AndresFPA/ModelSelection_Simulation* [GitHub project page containing code]. GitHub. https://github.com/AndresFPA/ModelSelection_Simulation

Perez Alonzo, A. F. (2025). *AndresFPA/mmgsem* [GitHub project page containing code and documentation]. GitHub. https://github.com/AndresFPA/ModelSelection_Simulation

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

Perez Alonso, A. F., Vermunt, J. K., Rosseel, Y. D., & De Roover, K.(2025). *Supplementary materials to* "Selecting the number of clusters in Mixture Multigroup Structural Equation Modeling" [Supplementary materials with simulation details and performance results]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.16173

R Core Team. (2024). *R: A language and environment for statistical computing*. R Project for Statistical Computing. https://www.R-project.org/

Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods, 29*(3), 561–588. https://doi.org/10.1037/met0000503

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: Recommendations and cautions. *Psychological Methods, 16*(1), 63–79. https://doi.org/10.1037/a0022673

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

van den Bergh, M., Schmittmann, V. D., & Vermunt, J. K. (2017). Building Latent Class Trees, with an application to a study of social capital. *Methodology, 13*(Supplement 1), 13–22. https://doi.org/10.1027/1614-2241/a000128

Vermunt, J. K., & Magidson, J. (2005, October.). Structural Equation Modeling: Mixture Models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. John Wiley & Sons. https://doi.org/10.1002/0470013192.bsa600

Vermunt, J. K., & Magidson, J. (2016). *Technical guide for Latent GOLD 5.1: Basic, advanced, and syntax*. Statistical Innovations.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17*(2), 228–243. https://doi.org/10.1037/a0027127