

## Mixture Multigroup Structural Equation Modeling for Ordinal Data


Andres F. Perez Alonso<sup>1, 2</sup>, Jeroen K. Vermunt<sup>1</sup>, Yves Rosseel<sup>3</sup>, and Kim De Roover<sup>2, 1</sup>


<sup>1</sup>Tilburg University


<sup>2</sup>KU Leuven


<sup>3</sup>Ghent University

### Author Note

Andres F. Perez Alonso  <https://orcid.org/0000-0002-2480-8771>

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

Yves Rosseel  <https://orcid.org/0000-0002-4129-4477>

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

This paper is still under review and has not been fully published yet. It is currently in its second version, following revisions from a peer-reviewed journal. Changes from its first version

have been highlighted in red. **Correspondence:** Correspondence concerning this article should be addressed to Andres F. Perez Alonso, Department of Methodology and Statistics, Tilburg University, PO Box 90153 5000 LE, Tilburg. E-mail: A.F.PerezAlonso@tilburguniversity.edu.

**Data and code availability:** The code behind the simulation, data generation, and analysis is openly available at <https://github.com/AndresFPA/OrdinalSim>.

**Acknowledgements:** 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). This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-11355.

**Disclosure of artificial intelligence-generated content (AIGC) tools:** No generative AI tools were used in the creation of this article.

### Abstract

Social scientists often compare groups in terms of relations between latent variables (LV) (often called structural relations) using Structural Equation Modeling (SEM). LVs are measured indirectly by questionnaires; thus, measurement invariance must be evaluated before comparisons can be made. To efficiently compare many groups, the recently proposed Mixture Multigroup SEM (MMG-SEM) clusters groups based on their structural relations while accounting for measurement (non-)invariance. However, the current MMG-SEM **relies on standard SEM implementations of maximum likelihood (ML) estimation in R**, which assume continuous indicators. This can introduce bias when dealing with ordinal data, especially for variables with fewer item categories. In this paper, we extend MMG-SEM to accommodate ordinal data relying on the stepwise Structural-After-Measurement estimation approach. In the first step, we implement a multigroup categorical confirmatory factor analysis (MG-CCFA) with diagonally weighted least squares (DWLS) to estimate the measurement model. The second step uses ML to perform the clustering and estimate cluster-specific structural relations. A simulation study compares the performance of this approach to that of ML-based MMG-SEM under various conditions. The results show a better recovery of measurement model parameters with DWLS, particularly with fewer response categories, whereas both approaches perform similarly regarding the recovery of the clusters and structural relations.

*Keywords:* ordinal data, mixture modeling, structural equation modeling, structural relations

## Mixture Multigroup Structural Equation Modeling for Ordinal Data

### 1 Introduction

Studying the relations between unobservable or ‘latent’ variables (e.g., emotions, ideology) is common in social sciences. Researchers often want to compare those relations, in the form of regression coefficients, across multiple groups. For instance, Abio et al. (2022) compared the effect of variables like anxiety and loneliness on suicide ideation across 33 countries. Structural Equation Modeling (SEM; Bollen, 1989) is the state-of-the-art for doing so, as it allows the estimation of relations between latent variables (LV) based on scores of observed indicators, often questionnaire items. In SEM, the LVs are called ‘factors’ and the regression coefficients are called ‘structural relations’.

When comparing structural relations across many groups, the two most common approaches within the SEM framework are Multigroup SEM (MG-SEM) and Multilevel SEM. Both approaches are well-established and valid methods for group comparison. However, they require pairwise comparisons of group-specific relations to pinpoint differences and similarities between specific groups, which can be a daunting task when many groups are involved. For example, for the 33 groups in Abio et al. (2022), 528 pairwise comparisons would be required.

Mixture modeling (McLachlan et al., 2019) can efficiently solve this problem by finding subsets or ‘clusters’ of groups that share the same relations. Before finding clusters of groups with the same structural relations, researchers must deal with the issue of measurement invariance. Since the LVs are indirectly measured by questionnaire items, it is necessary to ensure that the measurement of the factors is equivalent across groups. This equivalence is called ‘measurement invariance’ (MI; Meredith, 1993) and is a prerequisite for the comparability of structural relations. It involves evaluating whether the measurement model, indicating which items measure which factors (and to what extent), is the same across all groups. Note that MI can be evaluated at several levels, that is, regarding the equality of different parameters of the measurement model (Vandenberg & Lance, 2000).

When many groups are involved, some measurement parameters likely differ across groups (e.g., Davidov et al., 2014). Note that invariance of all measurement parameters is

not necessary for the comparability of the structural relations across groups (more details in the Method section). However, failing to account for measurement differences or ‘non-invariances’ (that is, forcing parameters to be equal across groups when they actually differ) can lead to incorrect estimates of the structural relations (Guenole & Brown, 2014), which are part of the so-called structural model. Thus, when looking for clusters of groups with the same structural relations, we should avoid these clusters being (partially) driven by measurement non-invariances.

To effectively find clusters of groups with equivalent structural relations while also accounting for measurement (non-)invariances, Perez Alonso et al. (2024) introduced Mixture Multigroup SEM (MMG-SEM). Building on the mixture multigroup approach introduced by De Roover (2021) and De Roover et al. (2022), MMG-SEM estimates cluster-specific structural relations, whereas measurement parameters are specific to each group so that the clustering is focused on the structural relations. This is in contrast to other mixture-based SEM methods (e.g., Kim et al., 2016; Vermunt & Magidson, 2005) that force all measurement model and structural model parameters to be cluster-specific (or equal across clusters) and thus capture clusters of groups with the same structural and measurement model. In elaborate simulation studies, MMG-SEM performed well in terms of recovering clusters of groups with the same structural relations (Perez Alonso et al., 2024), and model selection measures to determine the number of clusters (e.g., AIC, BIC) yielded promising results (Perez Alonso et al., 2025).

Up to now, MMG-SEM has been estimated via maximum likelihood (ML). ML is the traditional (or default) SEM estimation method, especially for mixture-based SEM methods, since they often use the expectation maximization (EM) algorithm (Dempster et al., 1977), which is specific to ML. In standard SEM software (e.g., *lavaan*; Rosseel, 2012), ML assumes the observed indicators (i.e., questionnaire items) to be continuous and typically normally distributed, whereas, in reality, ordinal variables are predominant in social sciences, where researchers often use Likert scales (e.g., ‘disagree’, ‘neutral’, ‘agree’). Several studies have evaluated the effect of ignoring the ordinal nature of observed variables in SEM under different conditions (Beauducel & Herzberg, 2006; Bollen, 1989; DiStefano, 2002; Dolan, 1994; Rhemtulla et al., 2012). Beauducel and Herzberg (2006),

Dolan (1994), and Johnson and Creech (1983) showed that the measurement model's parameter estimation is mostly unbiased when using ML for ordinal variables with five or more response categories, even when their distributions deviate from normality. In the case of items with less than five categories, Rhemtulla et al. (2012) showed that using ML estimation for SEM models led to substantially biased estimates of measurement parameters regardless of their distribution (but even more so in case of non-normality). This bias was reduced by using appropriate estimation methods, such as the widely recommended least squares (LS) estimator (Bollen, 1989; Flora & Curran, 2004; Kim & Yoon, 2011). Note that no bias was found in the structural parameters; that is, LS and ML correctly estimated the factor correlations.

Considering that ordinal variables with less than five categories are common in social sciences<sup>1</sup>, and the importance of avoiding biased estimates, extending MMG-SEM to accommodate ordinal variables is an important step and the aim of this paper. MMG-SEM is estimated in a step-wise fashion, following the 'Structural-After-Measurement' (SAM; Rosseel & Loh, 2022) approach. Specifically, MMG-SEM estimates the measurement model in Step 1 through multigroup confirmatory factor analysis (MG-CFA) and then models the relations between latent variables (and the corresponding clustering of the groups) in Step 2. This step-wise estimation allows for the use of different estimators in each step, making it possible to use an LS estimator for the measurement model in Step 1 while retaining the ML estimation of the cluster memberships and structural relations in Step 2 (using the EM algorithm). Specifically, we propose using multigroup categorical confirmatory factor analysis (MG-CCFA) to estimate the measurement model in MMG-SEM's first step. Categorical CFA is a well-established approach to model ordinal variables within the SEM framework. It assumes that there is an underlying continuous response behind the ordinal variables and uses the threshold model for its estimation (Muthén, 1984). Then, a standard factor analysis model is applied to the underlying continuous items. Note that SAM facilitates not only model estimation but also *model building*. The step-wise estimation conveniently allows applied researchers to address measurement model-related issues (e.g.,

---

<sup>1</sup> For instance, the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA).

misspecifications, MI, etc.) adequately before turning to their research interest (i.e., comparing structural relations). As discussed in Vermunt (2025), this helps prevent interpretational confounding and facilitates complying with standard practices (e.g., assessing model fit when doing a CFA).

Common software implementations of mixture SEM methods (e.g., Mplus; L. Muthen & Muthen, 2017) simultaneously estimate all SEM parameters and the cluster memberships via ML, making the implementation of LS estimation unfeasible. Of course, ML-based estimation of a SEM model with ordinal variables is possible, for instance, by modeling the relation between the items and the factors using a generalized linear model with an ordered probit link (Bartholomew et al., 2011; Rhemtulla et al., 2012), but such ML approaches require numerical integration (e.g., quadrature points), which becomes more computationally demanding as the sample size and number of the factors increase (L. Muthen & Muthen, 2017; Rhemtulla et al., 2012). Considering MMG-SEM aims to compare structural relations (between at least two factors, but often more) across many groups (i.e., large sample size), the step-wise estimation with LS in the first step is clearly the more time-efficient approach.

Note that the first step of MMG-SEM could also be estimated using Item Response Theory (IRT), which similarly relies on ML estimation with probit or logit links and is thus also computationally intensive for complex models (Forero & Maydeu-Olivares, 2009). Other reasons for not considering IRT in this paper are: (1) previous research has shown comparable results between IRT and CCFA when dealing with categorical data (Forero & Maydeu-Olivares, 2009; Kim & Yoon, 2011; Knol & Berger, 1991); (2) multigroup (partial) equality constraints for many groups are easier to implement in available software for CCFA than IRT; and (3) MMG-SEM was originally built within the SEM framework, using CFA rather than IRT.

The remainder of this paper is organized as follows: MMG-SEM and its estimation with a CCFA-based measurement model is described in the Method section. Then, a Simulation Study evaluates its performance and compares it to that of MMG-SEM using the standard ML estimator for the measurement model. **Next, we present an empirical application of MMG-SEM to data collected by the Programme for International Student**

Assessment (PISA) in 2018, following a similar analysis approach to that of OECD (2024).

The paper concludes with a Discussion section highlighting the study’s most relevant results and limitations.

## 2 Method

### 2.1 Mixture Multigroup Structural Equation Modeling

Mixture Multigroup Structural Equation Modeling (MMG-SEM) was proposed by Perez Alonso et al. (2024) as a flexible combination of mixture modeling and MG-SEM. In MMG-SEM, the structural relations are set as cluster-specific parameters, while the differences in the measurement model parameters are captured by group-specific parameters. Since the structural relations are the only cluster-specific parameters, MMG-SEM finds clusters based solely on the structural relations, which are usually the parameters of interest for the research question.

Perez Alonso et al. (2024) used the ‘Structural-After-Measurement’ (SAM; Rosseel & Loh, 2022) approach to estimate MMG-SEM. The SAM approach estimates SEM models in two steps. First, the measurement model is estimated, while the structural model is estimated in a second step. Below, we describe these two steps for MMG-SEM. In this paper, the first step (i.e., estimation of the multigroup measurement model) is adjusted to deal with the ordinal nature of the observed variables, whereas the second step (i.e., estimation of the structural model and clustering) stays the same. For a more detailed description of the second step, see the original paper from Perez Alonso et al. (2024).

### 2.2 Step 1: Measurement Model

The measurement model defines which items measure which LV and to what extent. For ordinal data from multiple groups, Multigroup Categorical CFA (MG-CCFA) is commonly used for its estimation. Consider data for items  $j = 1, \dots, J$  and individuals  $n_g = 1, \dots, N_g$  within groups  $g = 1, \dots, G$ . Then, for individual  $n_g$  let  $\mathbf{x}_{n_g}$  be a vector of ordinal observed polytomous scores, which can take  $c = 1, \dots, C$  possible values. MG-CCFA assumes there is an underlying latent vector of continuous responses  $\mathbf{x}_{n_g}^*$  that is discretized to obtain  $\mathbf{x}_{n_g}$  via a set of group- and item-specific threshold parameters  $\boldsymbol{\tau}_{jg}$ . Note that the  $\boldsymbol{\tau}_{jg}$  parameters are additional parameters in MG-CCFA compared to the

standard MG-CFA, for which  $\mathbf{x}_{n_g} = \mathbf{x}_{n_g}^*$ . For each item  $j$ , those thresholds divide the scores  $x_{j,n_g}^*$  into  $C$  observed categories as follows:

$$x_{j,n_g} = c \quad \text{if} \quad \tau_{j,c,g} < x_{j,n_g}^* < \tau_{j,c+1,g} \quad \text{for} \quad c = 1, \dots, C. \quad (1)$$

Considering factors  $q = 1, \dots, Q$ , MG-CCFA defines the vector of continuous scores  $\mathbf{x}_{n_g}^*$  of individual  $n_g$  as

$$\mathbf{x}_{n_g}^* = \mathbf{v}_g + \mathbf{\Lambda}_g \boldsymbol{\eta}_{n_g} + \boldsymbol{\epsilon}_{n_g}, \quad (2)$$

where  $\mathbf{v}_g$  is a  $J$ -dimensional vector of group-specific intercepts,  $\mathbf{\Lambda}_g$  denotes a  $J \times Q$  matrix of group-specific factor loadings (i.e., item-factor relations),  $\boldsymbol{\eta}_{n_g}$  is a  $Q$ -dimensional random vector of factor scores, and  $\boldsymbol{\epsilon}_{n_g}$  is a  $J$ -dimensional random vector of residuals. 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  $\text{Cov}(\boldsymbol{\eta}_{n_g}, \boldsymbol{\epsilon}_{n_g}) = \mathbf{0}$ , the model-implied covariance matrix of group  $g$  is given by

$$\boldsymbol{\Sigma}_g = \mathbf{\Lambda}_g \boldsymbol{\Phi}_g \mathbf{\Lambda}_g' + \boldsymbol{\Theta}_g. \quad (3)$$

### 2.2.1 Measurement Invariance

As mentioned in the Introduction, before comparing structural relations across groups, one must ensure that MI holds. MI can be evaluated at different levels in a step-wise fashion by focusing on different measurement model parameters (Svetina et al., 2020; Vandenberg & Lance, 2000). The first step is to test for configural invariance by checking if the model in Equation 2 holds across groups, which evaluates whether the same items measure the same factors for all groups. If the model fit is satisfactory (for details, see Chen, 2007), configural invariance holds. The second step is to evaluate threshold invariance by constraining the thresholds  $\boldsymbol{\tau}_g$  to be equal for all groups. It holds if the model fit does not significantly worsen (Rutkowski & Svetina, 2014) by imposing  $\boldsymbol{\tau}_g = \boldsymbol{\tau}$  for each group. The third step pertains to ‘metric’ invariance, which concerns the equality of



the factor loadings  $\Lambda_g$ . It holds if imposing  $\Lambda_g = \Lambda$  does not significantly lower the model fit. There are more MI levels (see Svetina et al., 2020; Vandenberg & Lance, 2000), but only threshold and metric invariance are a requisite for comparing structural relations (Wu & Estabrook, 2016). Therefore, by default, the remaining measurement model parameters (i.e.,  $\mathbf{v}_g$  and  $\Theta_g$ ) are allowed to vary per group in MMG-SEM. Given that ‘full’ threshold or metric invariance is difficult to attain, it is possible to aim for ‘partial’ invariance (e.g., a few loadings are free to vary per group; Byrne et al., 1989). If some non-invariant loadings and/or thresholds are identified, they should be specified as group-specific parameters in MMG-SEM.

In CCFA, the introduced latent continuous responses  $\mathbf{x}_{n_g}^*$  have unknown scales, which must be set through extra constraints. Several parameterizations are available to ensure model identification; the most common options being the delta (Christofferson, 1975) and theta parameterization (L. Muthen & Muthen, 2017). Under the delta parameterization, the variance of the observed variables is constrained to be exactly 1, whereas under the theta parameterization, the residual variances are constrained to be 1. In both cases, the thresholds  $\tau$  are freely estimated. When multiple groups are involved, a widely used identification rule was proposed by Millsap and Yun-Tein (2004), which fixes the variances of the observed variables to 1 in one reference group, while freely estimating them in the remaining groups. This is achieved by imposing additional constraints across groups (e.g., one loading per factor is fixed to 1, and some thresholds are invariant).

Although these parameterizations remain valid, we follow the recommendations from Wu and Estabrook (2016) when specifying models for evaluating different levels of MI. As they demonstrated, relying only on the delta parameterization or on the approach from Millsap and Yun-Tein (2004) may lead to issues when testing for invariance<sup>2</sup>, as different models can yield the same model fit but different parameter estimates depending on the imposed constraints. Wu and Estabrook (2016) partially uses the well-known ‘delta’ parameterization, but for our required level of MI, they only impose such constraints in the first group (i.e.,  $\text{diag}(\Sigma_1) = 1$ ). Additionally, they modify other identification constraints

---

<sup>2</sup> Although this paper does not evaluate MI testing, we still adopt the identification constraints recommended by Wu and Estabrook (2016) to be consistent with current best practices.

(e.g., on factor variances) based on the evaluated level of MI and the number of response categories.

To fit the measurement model, the difference between the model-implied covariance matrices  $\Sigma_g$  and the observed covariance matrix  $\mathbf{S}_g$  is minimized via Diagonally Weighted Least Squares (DWLS) estimation<sup>3</sup>. The group-specific factor covariance matrices  $\Phi_g$  from the first step are the input for MMG-SEM's second step. To avoid confusion with model-implied factor covariance matrices in Step 2, the covariance matrices  $\Phi_g$  from Step 1 will have a superscript  $s1$  (i.e.,  $\Phi_g^{s1}$ ) in what follows.

### 2.3 Step 2: Structural Model with Mixture Clustering

The second step pertains to the estimation of the factors' relations in the structural model and, in MMG-SEM, the clustering of the groups based on those relations. Since MMG-SEM clusters groups as a whole, all observations within a group are assigned to the same cluster. Due to the clustering, the structural model is conditional on the membership of group  $g$  to cluster  $k$ , which is denoted as  $z_{gk}$  and can take on a value of 0 or 1. As the true  $z_{gk}$  are unknown, one can only obtain the estimated  $\hat{z}_{gk}$ , which is a probability ranging from 0 to 1. Formally, the structural model defines the covariance matrix of the factors  $\Phi_{gk}$  as:

$$[\Phi_{gk} | z_{gk} = 1] = (\mathbf{I} - \mathbf{B}_k)^{-1} \Psi_{gk} (\mathbf{I} - \mathbf{B}_k)^{-1'}, \quad (4)$$

where  $\mathbf{B}_k$  is a non-symmetric  $Q \times Q$  matrix containing the unstandardized cluster-specific regression coefficients between LVs (structural relations), and  $\Psi_{gk}$  is the residual factor covariance matrix. The residual factor covariances  $\Psi_{gk}$  are specified as group-and-cluster-specific to ensure a clustering driven solely by the regression coefficients  $\mathbf{B}_k$ , and not by the residual factor covariances<sup>4,5</sup>.

The structural model is fitted via ML by minimizing the differences between the

---

<sup>3</sup> It is generally recommended to use Unweighted Least Squares or DWLS when dealing with ordinal data but there are no major differences between them in terms of parameter estimation (Yang-Wallentin et al., 2010). We used DWLS as it is the default estimator in the software we use (i.e., `lavaan`).

<sup>4</sup> For more details about the specification of  $\Psi_{gk}$ , please see the original paper by Perez Alonso et al. (2024).

<sup>5</sup> The group-cluster specification of  $\Psi_{gk}$  is incompatible with some identification constraints recommended by Wu and Estabrook (2016). For details on how this was solved, please see Appendix A.

model-implied factor covariance matrices  $\Phi_{gk}$  in Step 2 (Equation 4) and the group-specific covariance matrices  $\Phi_g^{s1}$  from Step 1. Note that the ordinal nature of the indicators was already accounted for in Step 1. Thus, we can safely rely on the default ML estimation for Step 2.

For the mixture clustering, MMG-SEM assumes the factor scores  $\eta_{n_g}$  are sampled from a mixture of  $K$  multivariate distributions where all  $\eta_{n_g}$  in group  $g$  — denoted by a matrix  $\mathbf{H}_g$  containing the factor scores — come from the same distribution. Formally, group  $g$  is defined as:

$$f(\mathbf{H}_g; \vartheta) = \sum_{k=1}^K \pi_k f_{gk}(\mathbf{H}_g; \vartheta_{gk}) = \sum_{k=1}^K \pi_k \prod_{n_g=1}^{N_g} MVN(\eta_{n_g}; \alpha_g, \Phi_{gk}), \quad (5)$$

where  $f$  is the population density function,  $\vartheta$  is the set of population parameters,  $\pi_k$  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 group  $g$  in the  $k$ th cluster, and  $\vartheta_{gk}$  is its corresponding set of parameters.

Specifically,  $f_{gk}$  is a multivariate normal distribution with  $\Phi_{gk}$  and  $\alpha_g$  as the covariance matrix and mean vector, respectively. Note that the elements in  $\alpha_g$  are fixed to 0 following the identification constraints in Wu and Estabrook (2016). The estimation of the unknown parameters  $\vartheta$  and the mixture clustering is done by maximizing the following log-likelihood function  $\log L_{\eta}$  via an Expectation Maximization algorithm (Dempster et al., 1977; Perez Alonso et al., 2024):

$$\log L_{\eta} = \sum_{g=1}^G \log \left( \sum_{k=1}^K \pi_k \left( \frac{1}{(2\pi)^{Q/2} |\Phi_{gk}|^{1/2}} \exp \left( -\frac{1}{2} \text{tr}(\Phi_g^{s1} \Phi_{gk}^{-1}) \right) \right)^{N_g} \right). \quad (6)$$

### 3 Simulation Study

#### 3.1 Design

The Simulation Study aimed to compare the performance of the two versions of MMG-SEM — that differ in terms of the estimation method in the first step (i.e., MG-CFA with ML versus MG-CCFA with DWLS) — when modeling ordinal data. Specifically, we used a Monte-Carlo simulation with seven manipulated factors to evaluate how well each version of MMG-SEM recovered the parameters and clustering. The following factors were

manipulated in a complete factorial design:

1. Size of regression parameters  $\beta$  (3 levels): 0.2, 0.3, 0.4;
2. Within-group sample size  $N_g$  (3 levels): 50, 100, 300;
3. Number of clusters  $K$  (2 levels): 2, 4;
4. Cluster size (2 levels): balanced, unbalanced;
5. Loadings non-invariance size (2 levels): 0.2, 0.4;
6. Thresholds non-invariance size (2 levels): 0.25, 0.50;
7. Number of categories of items (3 levels): 2, 4, 5

The number of groups was not manipulated, as previous similar simulation studies (Perez Alonso et al., 2025; Perez Alonso et al., 2024) have shown it to have little impact on the results. Instead, it was fixed at 36, which is within the generally found number of groups in empirical large-scale international surveys (Rutkowski & Svetina, 2014). The simulation design resulted in  $3$  (size of regression parameters)  $\times 3$  (within-group sample size)  $\times 2$  (number of clusters)  $\times 2$  (cluster size)  $\times 2$  (loadings non-invariance size)  $\times 2$  (threshold non-invariance size)  $\times 3$  (number of categories) = 432 data generation conditions. For each data generation condition, we generated 50 replications, for a total of  $432 \times 50 = 21600$  data sets. Moreover, for each version of MMG-SEM, two models were considered: one that correctly modeled the non-invariances (i.e., included them in the model; a partially invariant model) and one that disregarded the non-invariances (i.e., constrained *all* loading and threshold parameters to be equal across groups). Note that, for the version of MMG-SEM with continuous data, thresholds are not part of the measurement model; thus, for this version, the modeling of non-invariances concerns only the loadings. Therefore, each generated data set was analyzed four times, that is, with the two versions of MMG-SEM and with the non-invariances modeled or not. As a result, the Simulation Study included  $21600$  (data sets)  $\times 4 = 86400$  analyses in total.

It is important to note that, because MMG-SEM aims to compare the structural relations, we did not perform or evaluate MI testing in the simulation. We assumed

threshold and metric invariance to hold, and when non-invariances were modeled, we treated them as known rather than identifying them through MI testing<sup>6</sup>. For readers interested in MI testing with ordinal data, we refer to Svetina et al. (2020) for a comprehensive tutorial on this issue.

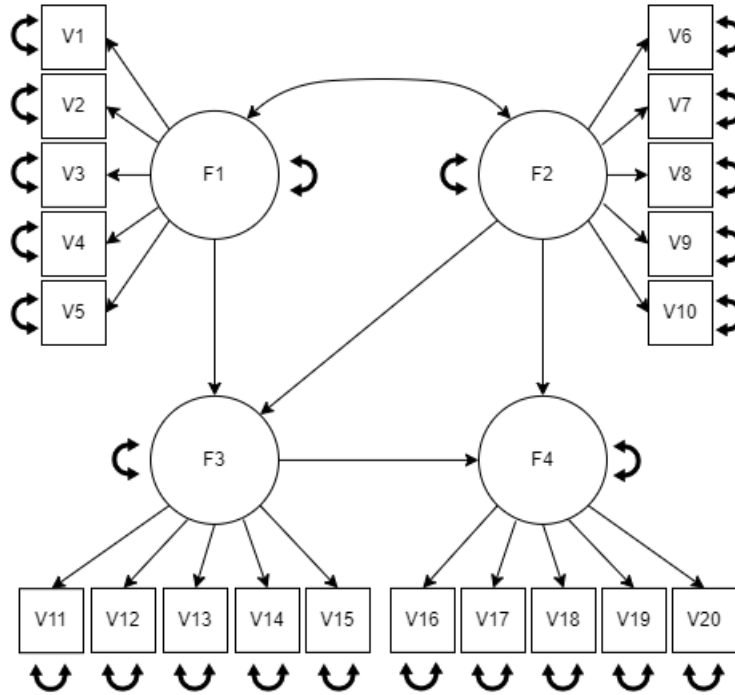
All the data generation and analysis of the results were done using R version 4.3.3 (RStudio Team, 2024). The code is openly shared on a GitHub repository that can be accessed at <https://github.com/AndresFPA/OrdinalSim>.

### 3.2 Data Generation

The data was generated following the SEM model depicted in Figure 1. The model includes four factors, defined as F1, F2, F3, and F4, measured by five indicators each, for a total of 20 observed variables. The structural relations (parameters of interest) are represented by the four regression parameters between the factors.

---

<sup>6</sup> Although the simulation study included dichotomous items, it is important to note that, in empirical research, it is not possible to test for threshold and metric invariance simultaneously for items with only two categories. In such a case, one can only assume that this level of invariance holds, as the corresponding model would be statistically equivalent, in terms of fit, to a configural model (for further details, see Wu & Estabrook, 2016).

**Figure 1**

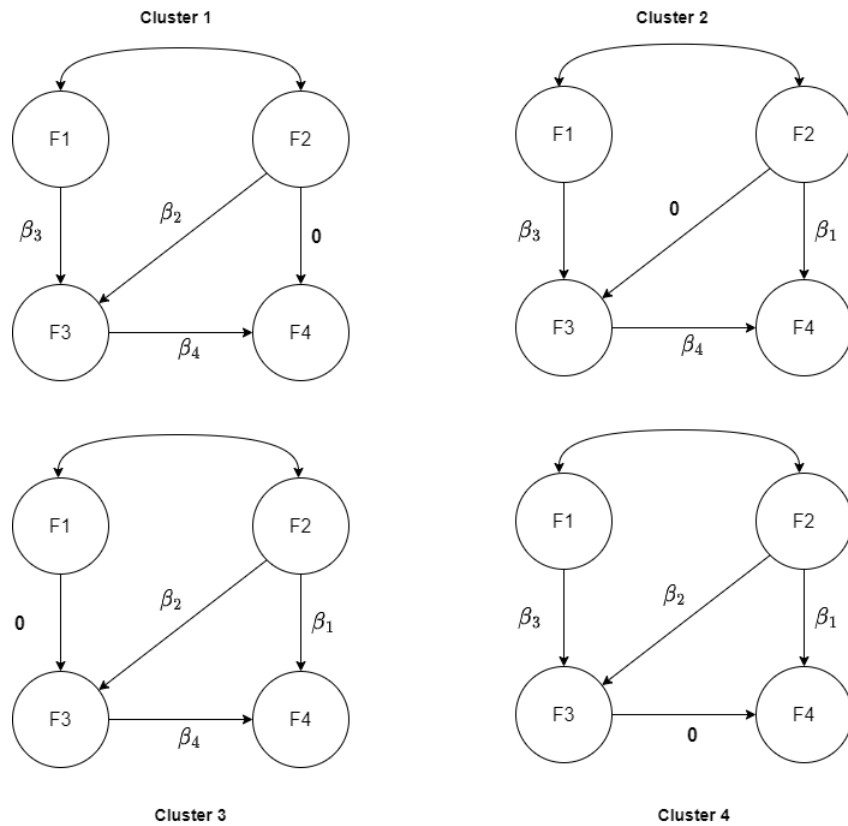
*The model used for the data generation.  $F1$  and  $F2$  are exogenous variables,  $F3$  is dependent and independent at the same time ('mediator'), and  $F4$  is only a dependent variable.*

Each generated data set pertained to 36 groups, and each group contained  $N_g$  observations following the manipulated factor 'within-group sample size' (i.e.,  $N_g$  equal to 50, 100 or 300). The number of clusters  $K$  was either 2 or 4 according to the manipulated factor 'number of clusters' and the number of groups per cluster depended on the manipulated factor 'cluster size'. When the cluster size was 'balanced', all clusters contained the same number of groups. For instance, when  $K = 4$ , each cluster contained  $36/4 = 9$  groups. In contrast, when cluster size was 'unbalanced', there was one larger cluster containing 75% of the groups, and the remaining (smaller) clusters had the same size. Thus, when  $K = 4$ , the first cluster contained 27 groups, and the remaining three clusters consisted of three groups each.

The cluster-specific structural relations were defined as depicted in Figure 2. The manipulated factor 'size of the regression parameters'  $\beta$  indicated the size of the

coefficients  $\beta_1$ ,  $\beta_2$ ,  $\beta_3$ , and  $\beta_4$ . Between-cluster differences in the regression coefficients were created by setting one of them to zero in each cluster. Thus, the size of the between-cluster differences depended on the size of the  $\beta$ . Note that, if  $K = 2$ , models three and four in Figure 2 were not applicable. The resulting cluster-specific regression parameters were gathered in the regression parameter matrix  $\mathbf{B}_k$ .

The factors' residual variances and covariances in the matrix  $\Psi_{gk}$  were obtained by sampling the variances  $\psi_{F1}$  and  $\psi_{F2}$  of factors F1 and F2 from a uniform distribution  $U(0.6, 0.8)$  and their covariance  $\psi_{F1,F2}$  from  $U(-0.3, 0.3)$  per group  $g$ . The total variances of factors F3 and F4 were also sampled from  $U(0.6, 0.8)$  for each group  $g$  and their residual variance depended on the cluster-specific regression coefficients. For instance, if the total variance of F3 for group  $g$  was  $\phi_{F3}$ , then the residual variance  $\psi_{F3}$  for the group-cluster combination  $gk$  was defined as  $\psi_{F3} = \phi_{F3} - (\beta_2^2\psi_{F1} + \beta_3^2\psi_{F2} + 2\beta_2\beta_3\psi_{F1,F2})$ .



**Figure 2**

*Zero and non-zero regression parameters between the factors depending on the cluster.*

To generate the observed ordinal variables, we followed the MG-CCFA model described in the Method section. That is, we first generated 20 underlying continuous variables from a  $MVN(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_{gk})$ , where the mean vector  $\boldsymbol{\mu}_g$  was a vector of zeros and the observed covariance matrix  $\boldsymbol{\Sigma}_{gk}$  was defined as

$$\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Lambda}(\mathbf{I} - \mathbf{B}_k)^{-1} \boldsymbol{\Psi}_{gk} (\mathbf{I} - \mathbf{B}_k)^{-1'} \boldsymbol{\Lambda}' + \boldsymbol{\Theta}_g, \quad (7)$$

by combining the measurement model from Equation 3 and the structural model of Equation 4. The parameters from the structural model (i.e.,  $\mathbf{B}_k, \boldsymbol{\Psi}_{gk}$ ) were generated as described above, and the measurement model parameters  $\boldsymbol{\Lambda}$  and  $\boldsymbol{\Theta}_g$  were generated aiming for a total variance of 1 per item, following the delta parameterization. To do this, the loadings in  $\boldsymbol{\Lambda}$  were set to  $\sqrt{0.5}$  for all groups, and the residual variances in the diagonal of  $\boldsymbol{\Theta}_g$  were defined for each item  $j$  in group  $g$ , as  $\theta_j = 1 - (\lambda_j^2 \psi_q)$ , where  $\psi_q$  is the corresponding factor's variance.

To evaluate the impact of loading non-invariances on MMG-SEM's performance, we added between-group differences to  $\boldsymbol{\Lambda}$ . Specifically, for all groups, we applied the non-invariances to each factor's second and third loading (the first loading is fixed to 1 to obtain a marker variable scaling). These non-invariances depended on the manipulated factor 'loadings non-invariance size' and were randomly sampled from a uniform distribution to ensure different values for each non-invariant group. For instance, if the size was 0.4, we sampled from  $U(0.3, 0.5)$ . Then, it was randomly decided if the non-invariance was added or subtracted from the original loading (i.e.,  $\sqrt{0.5}$ ). **Note that we ensured that the total item variance remained equal to 1, even for the items with non-invariant loadings.**

After generating a continuous vector  $\mathbf{x}_{j,g}^*$  for each item  $j$  per group  $g$ , the ordinal vector  $\mathbf{x}_{j,g}$  was obtained by discretizing  $\mathbf{x}_{j,g}^*$  into  $c$  categories using the thresholds  $\boldsymbol{\tau}_{j,g}$ . Initially, we generated invariant thresholds across groups (i.e.,  $\boldsymbol{\tau}_{j,g} = \boldsymbol{\tau}_j$ ). The categories  $c$  were defined by the manipulated factor 'number of item categories', and the thresholds  $\boldsymbol{\tau}_j$  were defined differently for each item. The first item was treated as an anchor item, and its thresholds were defined using quantiles to ensure that all categories contained at least one observation. For instance, if  $c = 2$ , then  $\tau_j$  was the median of  $\mathbf{x}_j^*$ . Similarly, if  $c = 4$  or



$c = 5$ , we used the quartiles or the quintiles, respectively. Given that it is empirically unrealistic to assume that all items have the same thresholds, the  $\tau_j$  for the remaining items were shifted randomly by -0.3 or 0.3 to ensure that not all items had the same threshold.

To evaluate the impact of threshold non-invariances on MMG-SEM’s performance, we added between-group differences to  $\tau_j$ . We followed a similar approach to the loadings non-invariance. We added a non-invariance to the second and third items of each factor following the manipulated factor ‘threshold non-invariance size’. The non-invariances were set to a quarter (0.25) and a half (0.5) standard deviation of the standardized  $\mathbf{x}_{j,g}^*$  to simulate small and large non-invariances, respectively (as in D’Urso et al., 2021). It was randomly decided whether these non-invariances were added or subtracted from the original threshold. Note that if the non-invariant item had more than one threshold (e.g., when  $c = 4$ , there are three thresholds per item), we added the non-invariance to all the thresholds of the item.

Lastly, it is worth mentioning that generating a ‘valid’ data set —that is, where the ordinal variables have at least one observation per category for each group— was difficult in some conditions, for instance, when  $c = 5$  and  $N_g = 50$ . Whenever a generated data set was not valid, we attempted to resample the data using the same data generation conditions but different random seeds. This procedure was repeated up to 50 times until a valid data set was found.

## 4 Results

To assess the performance of MMG-SEM, we evaluated the estimation of both the measurement model and the structural model. For the measurement model (Step 1 of MMG-SEM), we evaluated the recovery of the loadings<sup>7</sup>, whereas for the structural model (Step 2 of MMG-SEM), we evaluated the recovery of the clustering and the cluster-specific structural relations. We assessed the impact of ignoring and accounting for the ordinal

---

<sup>7</sup> As the loadings are the most important measurement model parameters affecting the recovery of the structural relations, we omitted studying the residual variances. Previously, Perez Alonso et al. (2024) found that residual variances had minimal impact on structural relations in MMG-SEM; and prior studies comparing LS and ML estimation have also focused on the assessment of the loadings (Dolan, 1994; Flora & Curran, 2004; Rhemtulla et al., 2012).

nature of the observed variables by comparing two versions of MMG-SEM, which differed in the estimation of Step 1. Specifically, one version used MG-CFA, and the other used MG-CCFA. To simplify the notation of the results, we refer to the results of both versions of MMG-SEM with a subscript corresponding to the estimator used in its first step (i.e., ML and DWLS, respectively). **Before discussing the findings, it is worth noting that disregarding the non-invariances (i.e., forcing them to be equal across groups) did not have a large impact on MMG-SEM’s performance.** Thus, for brevity, we describe only the results when the non-invariance is included. For more details about the results where we ignored the non-invariances, we refer the reader to Appendix B.

#### 4.1 Convergence rate

Before discussing the results, it is important to mention that MMG-SEM’s first step did not reach convergence for all data sets. Specifically, we found convergence problems for 1175 out of 21600 data sets (5.4%). For 732 of them, both MG-CFA and MG-CCFA models had convergence problems, whereas, for the remaining 433 data sets, only MG-CCFA failed to converge. When non-convergence was found, we ran the corresponding version of MMG-SEM one more time but with modified settings in its first step. Specifically, we relied on the bounded estimation (De Jonckere & Rosseel, 2022) implemented in the `lavaan` package, which has been shown to greatly reduce convergence problems in the presence of small sample sizes without compromising the parameter estimation. After rerunning both models for the mentioned data sets, we reached convergence for almost all of them. Convergence issues were found for only 299 data sets (1.3% of the total 21600). Of these, both estimators failed to converge for 259 data sets, while MG-CCFA and MG-CFA individually failed to converge for 25 and 15 data sets, respectively.

#### 4.2 Recovery of the factor loadings

We evaluated the recovery of the factor loadings using the Root Mean Squared Error (RMSE), which was computed as:

$$\text{RMSE}_\Lambda = \sqrt{\frac{\sum_{g=1}^G \sum_{j=1}^J (\lambda_{gj} - \hat{\lambda}_{gj})^2}{GJ}} \quad (8)$$

where  $\lambda_{gj}$  is the true factor loading of item  $j$  and  $\hat{\lambda}_{gj}$  is its corresponding estimate. The recovery was evaluated for both MG-CFA ( $\text{RMSE}_{\text{ML}}$ ) and MG-CCFA ( $\text{RMSE}_{\text{DWLS}}$ ). Note that we generated the data with the first loading of each factor equal to one (i.e., marker variable), whereas the estimation of MG-CCFA, following the guidelines from Wu and Estabrook (2016), does not use the marker variable approach. Thus, the estimations of all loadings were on a different scale. To correctly evaluate the loadings' recovery, we rescaled them before computing the RMSE. We excluded the first loading of each factor from the computation of  $\text{RMSE}_{\Lambda}$ , since they are fixed to one after the rescaling.

On average, the overall  $\text{RMSE}_{\Lambda, \text{ML}}$  and  $\text{RMSE}_{\Lambda, \text{DWLS}}$  were 0.118 and 0.105, respectively. The difference between the two estimators was more pronounced when the number of item categories was 2 ( $\text{RMSE}_{\Lambda, \text{ML}} = 0.213$ ;  $\text{RMSE}_{\Lambda, \text{DWLS}} = 0.118$ ) and decreased when the number of categories was 5 ( $\text{RMSE}_{\Lambda, \text{ML}} = 0.067$ ;  $\text{RMSE}_{\Lambda, \text{DWLS}} = 0.086$ ). The overall better performance of MG-CCFA compared to MG-CFA, especially when the number of item categories is lower than 5, is in line with previous research where the recovery of the measurement parameters was substantially worse when ignoring the ordinal nature of observed variables (Dolan, 1994; Johnson & Creech, 1983; Rhemtulla et al., 2012).

### 4.3 Cluster recovery

To evaluate whether MMG-SEM assigned the groups to the correct clusters, we used the Adjusted Rand Index (ARI; Hubert & Arabie, 1985). The ARI evaluates the similarity between two partitions (i.e., assignments of the groups to the clusters), and takes a value of 1 for complete agreement and 0 when the agreement is expected by pure chance. Thus, the ARI can take on negative values. Table 1 shows the ARI results and the proportion of perfect recovery or 'correct clustering' (CC; i.e., data sets for which the ARI was 1) for the main effects. Across all simulated conditions, the  $\text{ARI}_{\text{ML}}$  was 0.82 and the  $\text{CC}_{\text{ML}}$  was 0.588, whereas the  $\text{ARI}_{\text{DWLS}}$  was 0.837 and the  $\text{CC}_{\text{DWLS}}$  was 0.608. This shows that taking into account the ordinal nature of the variables led to a slightly better cluster recovery.

**Table 1***ARI and CC per level of each manipulated factor for both versions of MMG-SEM.*

Factor	Level	ARI <sub>ML</sub>	CC <sub>ML</sub>	ARI <sub>DWLS</sub>	CC <sub>DWLS</sub>
$\beta$ Size	0.2	0.642 (0.39)	0.352 (0.48)	0.675 (0.37)	0.383 (0.49)
	0.3	0.879 (0.22)	0.629 (0.48)	0.892 (0.21)	0.652 (0.48)
	0.4	0.939 (0.16)	0.781 (0.41)	0.943 (0.15)	0.788 (0.41)
$N_g$	50	0.600 (0.38)	0.294 (0.46)	0.638 (0.37)	0.320 (0.47)
	100	0.866 (0.22)	0.566 (0.49)	0.881 (0.20)	0.599 (0.49)
	300	0.988 (0.05)	0.895 (0.31)	0.987 (0.06)	0.897 (0.30)
$K$	2	0.876 (0.26)	0.699 (0.46)	0.893 (0.23)	0.721 (0.45)
	4	0.764 (0.33)	0.476 (0.49)	0.781 (0.32)	0.494 (0.50)
Cluster size	Bal	0.838 (0.29)	0.627 (0.48)	0.857 (0.27)	0.651 (0.48)
	Unbal	0.802 (0.31)	0.549 (0.49)	0.817 (0.29)	0.564 (0.49)
Non-inv Size $\lambda$	0.2	0.823 (0.30)	0.595 (0.49)	0.838 (0.28)	0.615 (0.49)
	0.4	0.817 (0.30)	0.580 (0.49)	0.835 (0.28)	0.600 (0.49)
Non-inv Size $\tau$	0.25	0.819 (0.30)	0.591 (0.49)	0.836 (0.29)	0.609 (0.49)
	0.50	0.820 (0.29)	0.584 (0.49)	0.837 (0.28)	0.606 (0.49)
Item categories $c$	2	0.690 (0.33)	0.342 (0.47)	0.721 (0.31)	0.375 (0.48)
	4	0.881 (0.26)	0.686 (0.46)	0.890 (0.24)	0.693 (0.46)
	5	0.888 (0.27)	0.735 (0.44)	0.900 (0.25)	0.755 (0.43)
Total	—	0.820 (0.30)	0.588 (0.49)	0.837 (0.28)	0.608 (0.49)

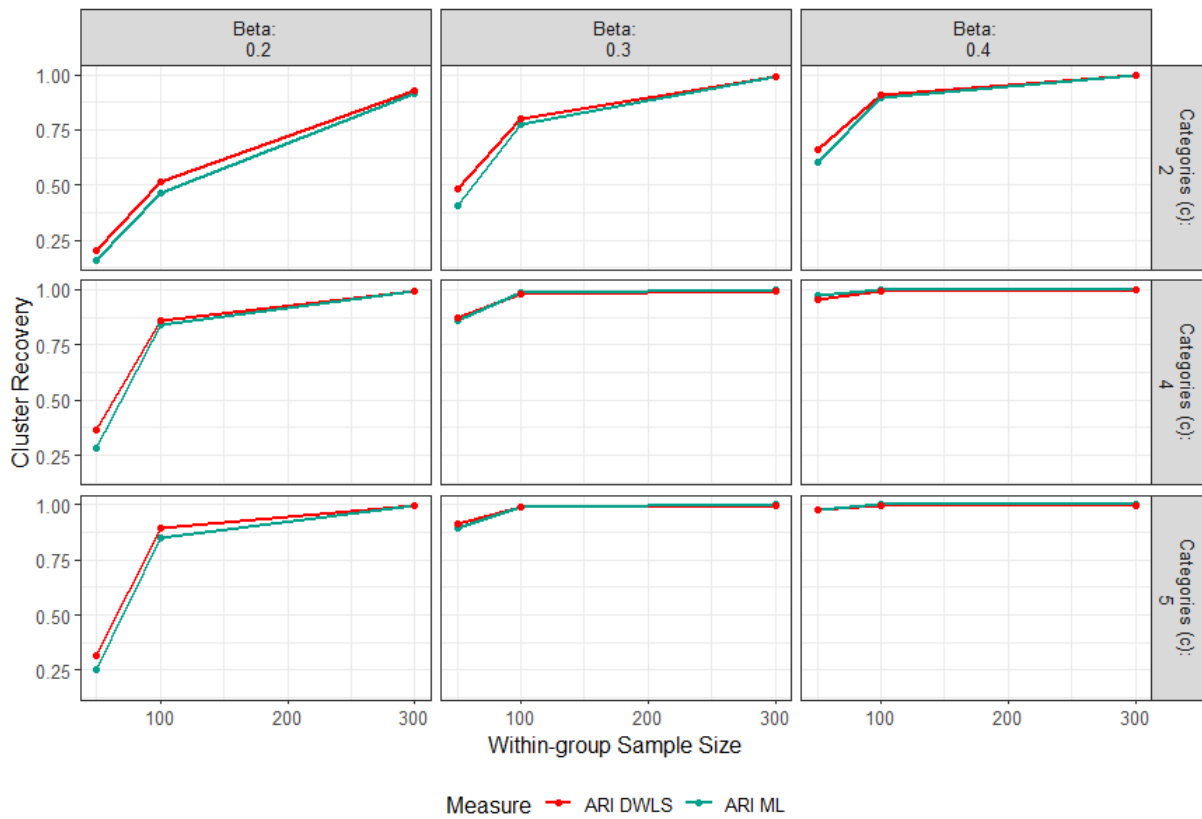
Note.  $\beta$  Size is the size of (difference in) regression parameters,  $N_g$  is the within-group sample size,  $K$  is the number of clusters, Non-inv Size  $\lambda$  is the size of the loadings' non-invariance, Non-inv Size  $\tau$  is the size of the thresholds' non-invariance, and  $c$  is the number of categories for each item.

The manipulated factors affected MMG-SEM similarly regardless of the estimator used in the first step. The factors that affected the performance the most were the sample size  $N_g$  and the size of the regression coefficient  $\beta$ , which aligns with the results obtained by Perez Alonso et al. (2024)<sup>8</sup>. Specifically, the larger the sample size, the better the cluster recovery. When  $N_g = 50$ , the ARI<sub>ML</sub> was 0.6, and the ARI<sub>DWLS</sub> was 0.638. This dramatically improved when  $N_g = 300$ , for which ARI<sub>ML</sub> was 0.988, and the ARI<sub>DWLS</sub> was 0.987. A similar trend can be seen for  $\beta$ , which determines the cluster separability. The ARI<sub>ML</sub> was 0.642 and 0.939, and the ARI<sub>DWLS</sub> was 0.675 and 0.943 when  $\beta$  was 0.2 and 0.4, respectively. We also found that the number of categories  $c$  of the items greatly affected the cluster recovery for both versions of MMG-SEM. The ARI<sub>ML</sub> was 0.69 and

<sup>8</sup> Note that the results in Perez Alonso et al. (2024) were generally slightly better, which may be partly due to differences in the simulation design and data generation procedure. Specifically, their data generation procedure involved sampling the group-specific data in such a way that its empirical covariance matrix matched the model-implied covariance matrix of the data generating model. In contrast, in our current simulation, we allow for the empirical covariance matrix to deviate from the data generating one due to sampling variability.

0.888, and the  $ARI_{DWLS}$  was 0.721 and 0.9 when  $c$  was 2 and 5, respectively.

The interaction between the most important factors can be seen in Figure 3. In the most difficult conditions (i.e.,  $N_g = 50$  and  $\beta = 0.2$ ), both versions of MMG-SEM failed to recover the original clustering. When  $N_g$  increased to 100 or  $\beta$  to 0.3, the ARI dramatically increased for both versions of MMG-SEM if  $c > 2$ . In contrast, if  $c = 2$ , both versions of MMG-SEM struggled to recover the correct clustering unless the cluster separation was high (i.e.,  $\beta = 0.4$ ) or the within-group sample size was large (i.e.,  $N_g = 300$ ).



**Figure 3**

*Cluster Recovery in function of Within-Group Sample Size, Size of the Regression Parameters, and Item Categories*

The remaining manipulated factors affected the performance of MMG-SEM to a lesser extent. Specifically, with more clusters, unbalanced cluster size, and/or larger non-invariances, the performance of both versions of MMG-SEM decreased.

#### 4.4 Regression Parameter Recovery

To evaluate the recovery of the regression parameters, we used the RMSE. It was defined as

$$\text{RMSE}_\beta = \sqrt{\frac{\sum_{k=1}^K (\beta_k - \hat{\beta}_k)^2}{K}} \quad (9)$$

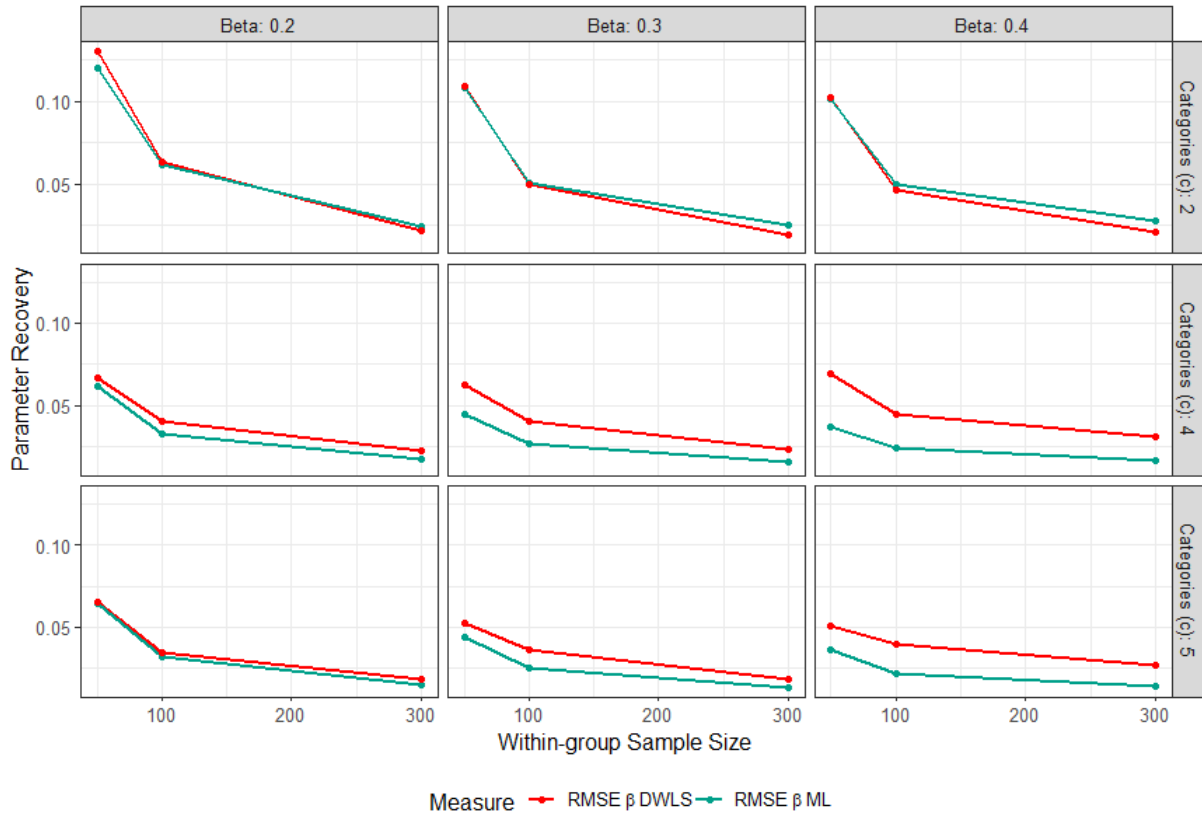
where  $K$  is the number of clusters,  $\beta_k$  is the true regression parameter, and  $\hat{\beta}_k$  is the estimated regression parameter. The main effects of the manipulated factors on the  $\text{RMSE}_\beta$  are shown in Table 2. The overall results are similar for both  $\text{RMSE}_{\beta, \text{ML}}$  (0.041) and  $\text{RMSE}_{\beta, \text{DWLS}}$  (0.048). Furthermore, the same trend from the cluster recovery results can be seen here when looking at the main effects. The most important manipulated factors were the within-group sample size  $N_g$ , the size of the regression parameters  $\beta$ , and the number of categories of the items  $c$ . Specifically,  $\text{RMSE}_{\beta, \text{ML}}$  was 0.068 and 0.019, and  $\text{RMSE}_{\beta, \text{DWLS}}$  was 0.079 and 0.022 when  $N_g$  was 50 and 300, respectively. Similarly, when  $\beta$  was 0.2 and 0.4,  $\text{RMSE}_{\beta, \text{ML}}$  was 0.048 and 0.036, and  $\text{RMSE}_{\beta, \text{DWLS}}$  was 0.050 and 0.048, respectively. Interestingly, MMG-SEM performed similarly ( $\text{RMSE} = 0.06$ ) if  $c = 2$ , regardless of the estimator used in the first step. In contrast, when  $c = 5$ , the performance of  $\text{RMSE}_{\beta, \text{ML}}$  (0.029) was slightly better than  $\text{RMSE}_{\beta, \text{DWLS}}$  (0.038).

**Table 2**RMSE $_{\beta}$  per level of each manipulated factor for both versions of MMG-SEM.

Factor	Level	RMSE $_{\beta,ML}$	RMSE $_{\beta,DWLS}$
$\beta$ Size	0.2	0.048 (0.047)	0.050 (0.054)
	0.3	0.039 (0.045)	0.046 (0.052)
	0.4	0.036 (0.041)	0.048 (0.060)
$N_g$	50	0.068 (0.059)	0.079 (0.066)
	100	0.036 (0.032)	0.044 (0.044)
	300	0.019 (0.014)	0.022 (0.035)
$K$	2	0.024 (0.023)	0.031 (0.036)
	4	0.058 (0.053)	0.065 (0.065)
Cluster size	Bal	0.032 (0.035)	0.039 (0.046)
	Unbal	0.050 (0.051)	0.057 (0.062)
Non-inv Size $\lambda$	0.2	0.040 (0.045)	0.045 (0.047)
	0.4	0.042 (0.044)	0.051 (0.062)
Non-inv Size $\tau$	0.25	0.041 (0.044)	0.048 (0.056)
	0.50	0.041 (0.045)	0.048 (0.054)
Item categories $c$	2	0.063 (0.060)	0.062 (0.066)
	4	0.031 (0.028)	0.044 (0.051)
	5	0.029 (0.029)	0.038 (0.042)
Total	—	0.041 (0.044)	0.048 (0.055)

Note.  $\beta$  Size is the size of (difference in) regression parameters,  $N_g$  is the within-group sample size,  $K$  is the number of clusters, Non-inv Size  $\lambda$  is the size of the loadings' non-invariance, Non-inv Size  $\tau$  is the size of the thresholds' non-invariance, and  $c$  is the number of categories for each item.

A deeper look into the interaction between these three manipulated factors is depicted in Figure 4. It is clear that RMSE $_{\beta,ML}$  performed slightly better than RMSE $_{\beta,DWLS}$  in most conditions. The performance difference between both versions of MMG-SEM decreased when the within-group sample size increased and was reversed when  $c = 2$ . Such results are in line with Rhemtulla et al. (2012), who found that, even when the observed variables are ordinal, the traditional ML estimation often performs slightly better than LS estimation in terms of structural parameter recovery as the number of categories  $c$  increases. This trend can be partially explained by the sample size, as studies such as Flora and Curran (2004) and Rhemtulla et al. (2012) showed that LS estimation requires a higher sample size ( $N_g > 200$ ) to produce unbiased structural parameters.

**Figure 4**

*Regression Parameter Recovery in function of Within-Group Sample Size, Size of the Regression Parameters, and Item Categories*

#### 4.5 Conclusion

Through a simulation study, we evaluated the performance of two versions of MMG-SEM: one that accounts for the ordinal nature of the observed variables in its measurement model (i.e., DWLS estimation) and one that does not (i.e., traditional ML estimation). We found that ignoring the ordinal nature of the variables greatly impacted the estimation of the measurement model parameters (i.e., loadings) but only had a minor influence on the estimation of the structural model. Indeed, the cluster and parameter recovery in the second step of MMG-SEM were remarkably similar for both versions. These results are in line with previous research, showing that ignoring the ordinal nature of variables mainly affect the measurement model estimation (DiStefano, 2002; Dolan, 1994; Rhemtulla et al., 2012). Notably, while ML estimation was generally slightly better than DWLS in recovering the structural relations, the DWLS estimation showed superior



performance in some scenarios, particularly when  $c = 2$  and  $N_g > 50$ , which may be due to the substantial bias in the loadings for ML estimation when  $c = 2$ .

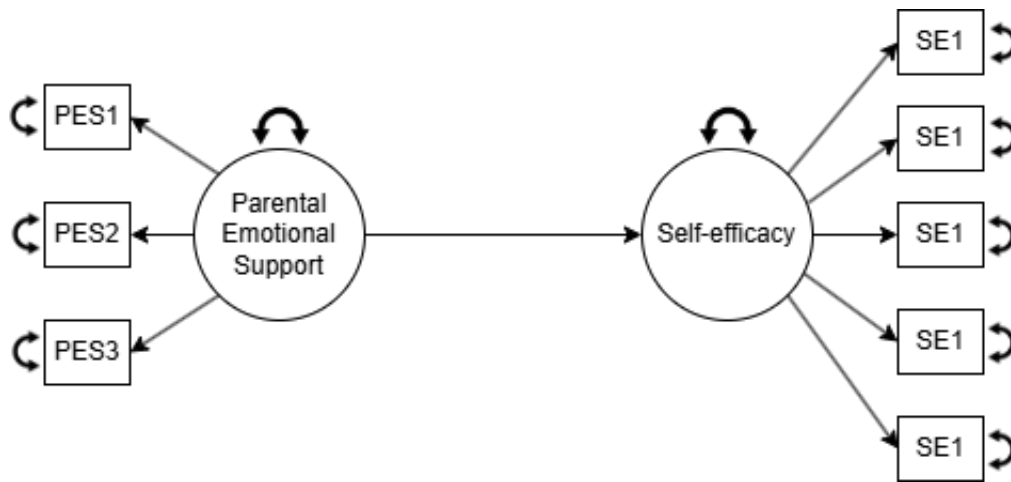
## 5 Empirical Application

In this section, we illustrate the empirical value of MMG-SEM for analyzing ordinal data using data collected by the PISA programme in 2018, conducted by the Organisation for Economic Co-operation and Development (OECD, 2019). PISA is a large-scale international survey administered every three years to assess the performance of young students (approximately 15 years old) in academic skills (i.e., mathematics, reading, science) as well as broader aspects of their lives, including family situation, school environment, and well-being, among others. The 2018 survey collected data from 612,005 students across 81 countries.

For the brief example presented in this paper, we followed part of the analysis reported in OECD (2024). Specifically, we focused on one of their research question: What is the effect of parental emotional support on the well-being of the students? In their study, the analysis was restricted to 29 OECD countries, and all cases with missing data were removed, resulting in a final sample of 169,961 participants. Parental emotional support was defined as the student's perception of the emotional and material resources provided by their parents. It was measured by three items on a 4-point Likert scale (1 = *strongly disagree*, 4 = *strongly agree*). Well-being was defined as students' thoughts and feelings about their lives, encompassing aspects such as a sense of purpose, perceptions of competence, and resilience, among others (Boarini et al., 2006). In OECD (2024), well-being was divided into seven subdomains (i.e., belonging to school, attitude towards competition, meaning in life, fear of failure, self-efficacy, positive feelings, and life satisfaction). For simplicity, this empirical example focused solely on self-efficacy, which was measured using five items on a 4-point Likert scale (1 = *strongly disagree*, 4 = *strongly agree*). Details about the items and questionnaires used can be found in OECD (2019, 2024).

In their original analysis, OECD (2024) used Item Response Theory (IRT) to create composite scores of their variables, which were then used in a linear regression model to examine the effect of parental emotional support on self-efficacy. While valid, this approach

has two main limitations. First, the authors did not evaluate measurement invariance<sup>9</sup>, which is essential for ensuring the validity of cross-country comparisons. Second, it is not easy to identify patterns across countries regarding their structural relations. With 29 groups, a total of 406 pairwise comparisons would be required to inspect all the differences and similarities in the structural relations of all countries. To address these issues, we re-analyzed the same data using MMG-SEM. The SEM model we used is shown in Figure 5, and an R script with the full analysis is available in the same GitHub repository as the simulation code: <https://github.com/AndresFPA/OrdinalSim>



**Figure 5**

*The SEM model depicting the relation between parental emotional support and self-efficacy. PES stands for parental emotional support, and SE for self-efficacy. The questionnaire items for both latent variables are described in OECD (2019).*

We analyzed the data with both versions of MMG-SEM: one using MG-CCFA and one using MG-CFA in the first step of its estimation. Because the results were broadly similar (and due to space constraints), we focus on the results using MG-CCFA in this section. The results using MG-CFA are available in Appendix C. Before running MMG-SEM, we conducted a measurement invariance test following the identification constraints proposed by Wu and Estabrook (2016) and the guidelines outlined by Svetina et al. (2020). Table 3 summarizes the results using scaled model fit indices. First, the configural model presented an acceptable fit ( $CFI = 0.989$  and  $RMSEA = 0.080$ ). Second,

<sup>9</sup> Differential item functioning, the IRT analogue to measurement invariance, was also not examined.

constraining the thresholds to equality across groups led to a decrease in fit ( $CFI = 0.987$  and  $RMSEA = 0.073$ ) that was within acceptable limits ( $<0.01$ ) (Rutkowski & Svetina, 2014). Finally, when the loadings were constrained to be equal across groups, the model fit ( $CFI = 0.986$ ,  $RMSEA = 0.069$ ) decreased minimally again, supporting the presence of metric invariance.

**Table 3**

*Model fit resulting from the measurement invariance testing per level.*

Measurement Invariance level	CFI	RMSEA
Configural	0.989	0.080
Threshold	0.987	0.073
Metric	0.986	0.069

After satisfactorily establishing measurement invariance, we proceeded to analyze the structural relations using MMG-SEM. As is common in empirical applications, the ‘true’ number of clusters was unknown. Following the recommendations of Perez Alonso et al. (2025), we selected the optimal number of clusters using  $AIC_3$  (Bozdogan, 1994)—a variation of the Akaike Information Criteria (AIC; Akaike, 1974)—and the Convex Hull scree ratio measure (Ceulemans & Kiers, 2006). We compared models with 1 to 6 clusters, and both measures selected the 3-cluster model (see the GitHub repository for additional details)<sup>10</sup>.

The cluster assignments are shown in Table 4. While Cluster 1 contained a single isolated country (i.e., South Korea), Clusters 2 and 3 presented some interesting geographical grouping. Cluster 2 primarily consisted of countries from Western Europe, with additional representation from Chile, Japan, and a few Central and Eastern European countries (e.g., Latvia, Poland). In contrast, Cluster 3 primarily consisted of countries from North and Eastern Europe, as well as those from the Americas (i.e., the USA, Mexico, and Colombia).

<sup>10</sup> Interestingly, the two versions of MMG-SEM diverged in model selection: with MG-CCFA, the 3-cluster model was preferred, whereas with MG-CFA, the 2-cluster model was selected (see Appendix C for details).

**Table 4**

*Clustering of the countries based on the regression parameter between parental emotional support and self-efficacy (3-cluster model).*

Cluster	Countries
Cluster 1	South Korea
Cluster 2	Austria, Switzerland, Chile, Germany, Spain, France, United Kingdom, Ireland, Iceland, Japan, Latvia, Netherlands, Poland, Portugal, Slovakia
Cluster 3	Colombia, Czech Republic, Estonia, Finland, Greece, Hungary, Lithuania, Luxembourg, Mexico, Slovenia, Sweden, Turkey, United States of America

The resulting cluster-specific regression parameters are presented in Table 5.

Although the size of the effect varied, the relation between parental emotional support and self-efficacy was positive for all three clusters, indicating that higher parental support predicts higher self-efficacy across all 29 countries. Regarding between-cluster differences, South Korea clearly showed the largest regression coefficient ( $\beta = 0.469$ ) compared to any other country. In comparison, Cluster 2, comprising mainly Western European countries, presented the smallest effect ( $\beta = 0.254$ ), while Cluster 3 had the second-largest regression coefficient ( $\beta = 0.335$ ).

**Table 5**

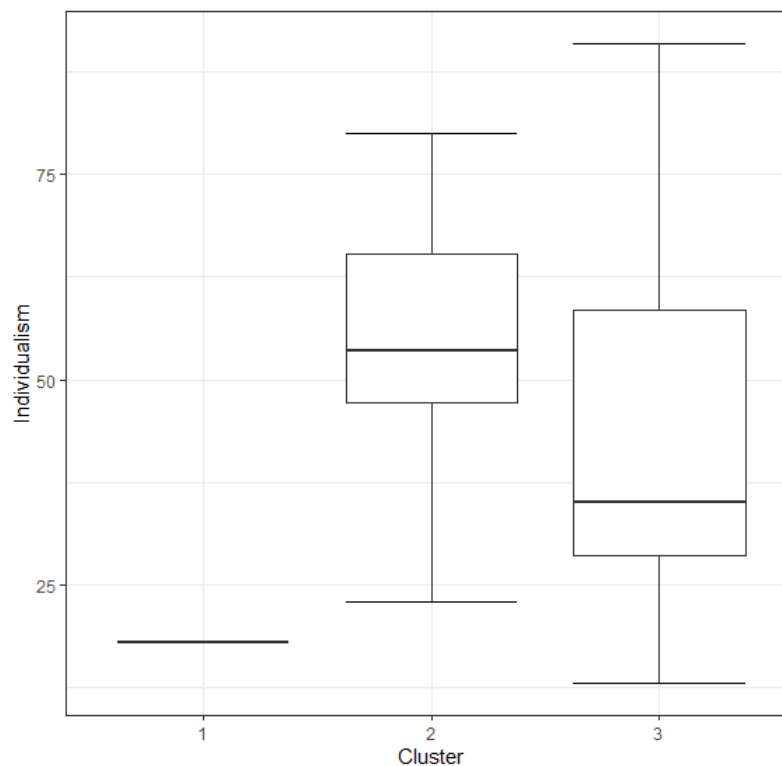
*Regression parameters per cluster for the relation between parental emotional support and self-efficacy.*

	Cluster 1	Cluster 2	Cluster 3
Regression parameter	0.469	0.254	0.335

Previous research has also highlighted cross-national differences in the relation between parental support and self-efficacy. For instance, Youn et al. (2023) found a stronger effect in South Korea compared to the USA, which aligns with our results, where South Korea had the largest effect compared to other countries. However, while research has examined moderators of this relation *within* countries (e.g., gender, parental education; OECD, 2024), relatively little is known about the country-level variables that may explain cross-national variation. Potential variables include cultural factors such as collectivism/individualism (Marbell-Pierre et al., 2019), characteristics of the education systems (Parveva et al., 2020), or income inequalities (Qian et al., 2024).

To illustrate this, we conducted a brief explorative comparison of country-level

individualism scores across clusters (Figure 6). We used the scores from Hofstede (2001), which had data for 18 out of the total 29 countries of the empirical example. The scores ranged from 0 to 100, with higher scores indicating more individualistic countries. The boxplot clearly shows mean differences in individualism scores: Cluster 2 had the largest mean individualism score, followed by Cluster 3, and then Cluster 1. Interestingly, this order was the reverse of the regression coefficients per cluster. This may suggest that more individualistic societies display weaker relations between parental emotional support and self-efficacy. However, these results must be interpreted with caution as, for instance, Cluster 3 also contained both the country with the highest (i.e., USA) and lowest (i.e., Colombia) individualism scores. Thus, there may be more variables moderating this relation across countries.



**Figure 6**  
*Individualism scores for the three clusters obtained by MMG-SEM.*

In conclusion, MMG-SEM revealed meaningful cross-country differences in the relation between parental emotional support and self-efficacy. By clustering countries, MMG-SEM simplified the complexity of many-groups comparisons, resulting in three

clusters with distinct regression parameters. We also provided an exploratory analysis suggesting that cultural factors, such as individualism, may explain the cross-country variations. Finally, although the results were broadly similar for both versions of MMG-SEM, it is important to note that correctly treating data as ordinal and properly testing for measurement invariance are crucial steps to ensure the validity of cross-country comparisons of the structural relations, which are the main parameters of interest in this kind of research question.

## 6 Discussion

Social scientists often work with ordinal data, such as Likert scales, which pose challenges for conventional ML estimation methods in SEM. When using standard multiple-group methods like MG-SEM to compare structural relations, researchers can use DWLS estimation to deal with ordinal data correctly. However, if one wants to use mixture SEM methods (Kim et al., 2016; Vermunt & Magidson, 2005) to compare many groups, using DWLS estimation is difficult. Mixture SEM methods usually estimate all SEM parameters and the cluster membership simultaneously via ML, which is very time consuming when taking the ordinal nature of the observed variables into account. In this paper, we extended the recently developed MMG-SEM (Perez Alonso et al., 2024) —for comparing structural relations across many groups— to deal with ordinal data through a combination of DWLS and ML estimation. Specifically, thanks to the stepwise estimation of MMG-SEM, it was possible to implement MG-CCFA with DWLS estimation in the first step (effectively dealing with ordinal data) while maintaining the ML estimation for the structural model and clustering in the second step (using the EM algorithm).

We provided an empirical example of properly treating data as ordinal when using MMG-SEM, and performed an extensive simulation study to evaluate how the new version of MMG-SEM (using MG-CCFA in the first step) compared to the previous version of MMG-SEM (using MG-CFA in the first step) when analyzing ordinal data. In addition, we examined the impact of correctly modeling or disregarding non-invariances on model estimation. Since disregarding the non-invariances had only minor effects on the estimates, our analyses focused on the models that correctly captured the non-invariances, which can be viewed as partially invariant models. Overall, appropriately dealing with ordinal

variables improved MMG-SEM's performance in terms of measurement model parameter recovery. Specifically, the factor loading estimates were considerably more biased when using MG-CFA instead of MG-CCFA, especially in case of fewer item categories (i.e.,  $c = 2$ ). In contrast, using DWLS or ML in MMG-SEM's first step did not significantly affect the results of the second step (i.e., the recovery of the clustering and cluster-specific structural relations), despite building on the estimated factor loadings, which is line with previous research (Rhemtulla et al., 2012). Therefore, it might be tempting to always use MG-CFA (i.e., ML estimation) in the first step of MMG-SEM, regardless of the nature of the observed variables. However, to validly evaluate measurement invariance, we recommend using MG-CCFA (i.e., DWLS estimation) in MMG-SEM's first step, as it obtains more accurate measurement model parameters if the observed variables are ordinal, which is especially important when the number of item categories is low. Also, when  $c = 2$ , the cluster and regression parameter recovery were slightly worse when ignoring the ordinal nature of the variables.

The optimistic results should be interpreted in light of the limitations of the simulation study. Specifically, we cannot generalize the obtained results outside the conditions evaluated in the simulation. For simplicity, we generated data with underlying normal distributions and symmetrically distributed ordinal responses, whereas empirical data can present a wide variety of distributions. Previous research shows that non-normality can substantially increase bias in loading estimates under ML estimation, especially with few response categories (Dolan, 1994; Mîndrilă, 2010; Rhemtulla et al., 2012), whereas LS estimators are more robust to such deviations. However, these findings are based on simple CFA models, and it remains important to examine how non-normality affects MMG-SEM. In this study, we evaluated DWLS in the first step and did not consider ML-based models for ordinal variables (i.e., CFA with probit links or IRT) due to their higher computational demands and comparable performance to LS estimators (Forero & Maydeu-Olivares, 2009; Kim & Yoon, 2011; Liang & Yang, 2014). Nonetheless, future research can explore how such estimators perform when dealing with ordinal data in the context of MMG-SEM.

**It should also be noted that our simulation assumed MI and did not test it. As a**

result, we avoided the challenges inherent in invariance testing, such as selecting the anchor/marker variable and identifying the non-invariant parameters, which are necessary steps to ensure the validity of cross-group comparisons. In our simulation, the non-invariant parameters of the partially invariant models were treated as known. For an in-depth tutorial on traditional MI testing with ordinal data, we refer the readers to Svetina et al. (2020).

Moreover, although partial measurement invariance is commonly used to relax the strict requirements of full invariance, which is difficult to achieve when many groups are involved (e.g., Davidov et al., 2014), it remains a topic of debate in the SEM literature. For instance, Shi et al. (2019) showed that structural parameters may be unbiased even when only a single loading (i.e., the marker variable loading) is invariant. However, as they noted, when most of the loadings are non-invariant, it becomes difficult to argue that they measure the same construct. Alternative approaches, such as Bayesian CFA for ordinal data (Liang & Yang, 2014; Muthén & Asparouhov, 2012), provide ways to evaluate MI without imposing full (or partial) exact invariance, which would be interesting to explore in the future. Specifically, Bayesian CFA allows for approximate invariance (B. Muthen & Asparouhov, 2013), in which parameters are constrained to be *approximately*, rather than *exactly*, equal across groups.



## References

- Abio, A., Owusu, P. N., Posti, J. P., Bärnighausen, T., Shaikh, M. A., Shankar, V., & Lowery Wilson, M. (2022). Cross-national examination of adolescent suicidal behavior: A pooled and multi-level analysis of 193,484 students from 53 LMIC countries. *Social Psychiatry and Psychiatric Epidemiology*, 57(8), 1603–1613. <https://doi.org/10.1007/s00127-022-02287-x>
- 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>
- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. John Wiley & Sons, Incorporated. Retrieved January 27, 2025, from <http://ebookcentral.proquest.com/lib/uvtilburg-ebooks/detail.action?docID=819225>
- Beauducel, A., & Herzberg, P. Y. (2006). On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186–203. [https://doi.org/10.1207/s15328007sem1302\\_2](https://doi.org/10.1207/s15328007sem1302_2)
- Boarini, R., Johansson, Å., & Mira d’Ercole, M. (2006). Alternative Measures of Well-Being [Series: OECD Social, Employment and Migration Working Papers Volume: 33]. *OECD Social, Employment and Migration Working Papers*, 33. <https://doi.org/10.1787/713222332167>
- 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](https://doi.org/10.1007/978-94-011-0800-3_3)
- 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>
- 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>
- Christofferson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, 40(1), 5–32. <https://doi.org/10.1007/BF02291477>
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, 40(1), 55–75.  
<https://doi.org/10.1146/annurev-soc-071913-043137>
- De Jonckere, J., & Rosseel, Y. (2022). Using Bounded Estimation to Avoid Nonconvergence in Small Sample Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(3), 412–427.  
<https://doi.org/10.1080/10705511.2021.1982716>
- 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>
- 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>

- DiStefano, C. (2002). The Impact of Categorization With Confirmatory Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(3), 327–346.  
[https://doi.org/10.1207/S15328007SEM0903\\_2](https://doi.org/10.1207/S15328007SEM0903_2)
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data [\_eprint: <https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2044-8317.1994.tb01039.x>]. *British Journal of Mathematical and Statistical Psychology*, 47(2), 309–326. <https://doi.org/10.1111/j.2044-8317.1994.tb01039.x>
- D’Urso, E. D., De Roover, K., Vermunt, J. K., & Tijmstra, J. (2021). Scale length does matter: Recommendations for measurement invariance testing with categorical factor analysis and item response theory approaches. *Behavior Research Methods*, 54(5), 2114–2145. <https://doi.org/10.3758/s13428-021-01690-7>
- Flora, D. B., & Curran, P. J. (2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis With Ordinal Data. *Psychological Methods*, 9(4), 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14(3), 275–299. <https://doi.org/10.1037/a0015825>
- Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00980>
- Hofstede, G. (2001). *Culture’s consequences Comparing values, behaviors, institutions, and organizations across nations* (2nd). Sage Publications.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218. <https://doi.org/10.1007/BF01908075>
- Johnson, D. R., & Creech, J. C. (1983). Ordinal Measures in Multiple Indicator Models: A Simulation Study of Categorization Error. *American Sociological Review*, 48(3), 398–407. <https://doi.org/10.2307/2095231>

- Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2024). *semTools: Useful tools for structural equation modeling*.  
<https://CRAN.R-project.org/package=semTools>
- 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>
- Kim, E. S., & Yoon, M. (2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling: A Multidisciplinary Journal*, 18(2), 212–228.  
<https://doi.org/10.1080/10705511.2011.557337>
- Knol, D. L., & Berger, M. P. (1991). Empirical Comparison Between Factor Analysis and Multidimensional Item Response Models. *Multivariate Behavioral Research*, 26(3), 457–477. [https://doi.org/10.1207/s15327906mbr2603\\_5](https://doi.org/10.1207/s15327906mbr2603_5)
- Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, 2(1), 17.  
<https://doi.org/10.1504/IJQRE.2014.060972>
- Marbell-Pierre, K. N., Grolnick, W. S., Stewart, A. L., & Raftery-Helmer, J. N. (2019). Parental Autonomy Support in Two Cultures: The Moderating Effects of Adolescents' Self-Construals. *Child Development*, 90(3), 825–845.  
<https://doi.org/10.1111/cdev.12947>
- 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>
- Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures [Publisher: Routledge \_eprint:

- [https://doi.org/10.1207/S15327906MBR3903\\_4](https://doi.org/10.1207/S15327906MBR3903_4). *Multivariate Behavioral Research*, 39(3), 479–515. [https://doi.org/10.1207/S15327906MBR3903\\_4](https://doi.org/10.1207/S15327906MBR3903_4)
- Mîndrilă, D. (2010). Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) Estimation Procedures: A Comparison of Estimation Bias with Ordinal and Multivariate Non-Normal Data. *International Journal for Digital Society*, 1(1), 60–66. <https://doi.org/10.20533/ijds.2040.2570.2010.0010>
- Muthen, B., & Asparouhov, T. (2013). BSEM Measurement Invariance Analysis. <http://www.statmodel.com/examples/webnotes/webnote17.pdf>
- Muthen, L., & Muthen, B. (2017). *Mplus User's Guide* (Eight Edition). Muthen & Muthen.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1), 115–132. <https://doi.org/10.1007/BF02294210>
- Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. [Publisher: American Psychological Association]. *Psychological Methods*, 17(3), 313–335. <https://doi.org/10.1037/a0026802>
- OECD. (2019, April). *PISA 2018 Assessment and Analytical Framework*. <https://doi.org/10.1787/b25efab8-en>
- OECD. (2024). Parental emotional support and adolescent well-being: A cross-national examination of socio-economic and gender gaps based on PISA 2018 surveys. *OECD Papers on Well-being and Inequalities*, 20. <https://doi.org/https://doi.org/10.1787/2b7a2ac6-en>
- Parveva, T., Horváth, A., Krémó, A., & Sigalas, E. (2020). *Equity in school education in Europe: Structures, policies and student performance*. Publications Office of the European Union. Retrieved September 30, 2025, from <https://data.europa.eu/doi/10.2797/658266>
- Perez Alonso, A. F., Vermunt, J. K., Rosseel, Y., & Roover, K. D. (2025). Selecting the number of clusters in Mixture Multigroup Structural Equation Modeling. *Methodology*, 21(1), 1–26. <https://doi.org/10.5964/meth.14931>

- 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>
- Qian, M., Jin, R., Lu, C., & Zhao, M. (2024). Parental emotional support, self-efficacy, and mental health problems among adolescents in Hong Kong: A moderated mediation approach [Publisher: Frontiers]. *Frontiers in Psychiatry*, 15.  
<https://doi.org/10.3389/fpsy.2024.1458275>
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, 17(3), 354–373. <https://doi.org/10.1037/a0029315>
- Rosseel, Y. (2012). **lavaan** : An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Rosseel, Y., & Loh, W. W. (2022). A structural after measurement approach to structural equation modeling. *Psychological Methods*. <https://doi.org/10.1037/met0000503>
- RStudio Team. (2024). RStudio: Integrated Development Environment for R.  
<http://www.rstudio.com/>
- 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>
- Shi, D., Song, H., & Lewis, M. D. (2019). The Impact of Partial Factorial Invariance on Cross-Group Comparisons [Publisher: SAGE Publications Inc]. *Assessment*, 26(7), 1217–1233. <https://doi.org/10.1177/1073191117711020>
- Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using.
- 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>

- Vermunt, J. K. (2025). Stepwise estimation of latent variable models: An overview of approaches [Publisher: SAGE Publications India]. *Statistical Modelling*, 1471082X251355693. <https://doi.org/10.1177/1471082X251355693>
- 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, Ltd.  
<https://doi.org/10.1002/0470013192.bsa600>
- Wu, H., & Estabrook, R. (2016). Identification of Confirmatory Factor Analysis Models of Different Levels of Invariance for Ordered Categorical Outcomes. *Psychometrika*, 81(4), 1014–1045. <https://doi.org/10.1007/s11336-016-9506-0>
- Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory Factor Analysis of Ordinal Variables With Misspecified Models [Publisher: Routledge \_eprint: <https://www.tandfonline.com/doi/pdf/10.1080/10705511.2010.489003>]. *Structural Equation Modeling: A Multidisciplinary Journal*, 17(3), 392–423.  
<https://doi.org/10.1080/10705511.2010.489003>
- Youn, J., Napolitano, C. M., Han, D., Lee, W., & Rounds, J. (2023). A meta-analysis of the relations between parental support and children’s career self-efficacy in South Korea and the US. *Journal of Vocational Behavior*, 141, 103839.  
<https://doi.org/10.1016/j.jvb.2022.103839>

## Appendix A

### Rescaling the standardized factor variances

In this paper, we adopted the parameterization approach for categorical CFA models with ordinal data as proposed by Wu and Estabrook (2016) for MMG-SEM’s first step. For its implementation, we used the `semTools` R package (Jorgensen et al., 2024). This approach sets the scale of the latent variables by standardizing them—i.e., setting their variances to one. In contrast, the original implementation of the first step in MMG-SEM (i.e., a standard MG-CFA; Perez Alonso et al., 2024) used the marker variable approach (i.e., one loading per factor, by default the first one, is fixed to one), which implies unstandardized factor covariances. As MMG-SEM makes use of the group-and-cluster-specific specification of the residual factor covariances (Equation 4) in its second step, it was not appropriate to directly use the standardized factor covariances resulting from the parametrization from Wu and Estabrook (2016). To solve this problem, the group-specific standardized factor covariances obtained in Step 1 were rescaled to correspond to the marker variable scale before feeding them to Step 2.

Formally, the rescaling procedure was defined as follows:

$$\phi_q^{uns} = \phi_q^{std} \lambda_{iq}^{mkr^2} \quad (\text{A1})$$

where  $q$  is a latent factor,  $\phi_q^{std}$  is factor  $q$ ’s *standardized* factor variance resulting from the parametrization from Wu and Estabrook (2016),  $\phi_q^{uns}$  is the *unstandardized* version of the factor variance, and  $\lambda_{iq}^{mkr}$  is the factor loading of the ‘marker’ variable  $i$  related to factor  $q$ .



## Appendix B

### Simulation Results - Ignoring Measurement Non-invariance

This Appendix briefly outlines the results of the simulation study when the measurement non-invariances were ignored. Specifically, we ignored the loading non-invariances when using MG-CFA and ignored the loading and threshold non-invariances when using MG-CCFA (i.e., these parameters were forced to be equal across groups). The overall results when ignoring the non-invariances are similar to the main text, especially regarding the structural model. However, there are differences in performance, which will be described below. Note that, as in the main text, we use a subscript corresponding to the estimator used in its first step (i.e., ML and DWLS) to distinguish between the two versions of MMG-SEM.

#### 6.1 Recovery of the factor loadings

We evaluated the loading recovery using RMSE (Equation 8). The average  $RMSE_{\Lambda, ML}$  and  $RMSE_{\Lambda, DWLS}$  were 0.22 and 0.24, respectively, indicating a notable decline in performance for both estimators compared to when the non-invariances were modeled (see Section 4.2). Interestingly, the ML estimator now outperforms the DWLS estimator. This outcome can be explained by the fact that, in MG-CFA, only loading non-invariances are ignored, whereas in MG-CCFA both loading and threshold non-invariances are disregarded, causing a greater negative impact on the performance of the DWLS estimator.

#### 6.2 Cluster Recovery

Table B1 shows the ARI and CC results for the main effects of all manipulated factors. Across all simulated conditions, the  $ARI_{ML}$  was 0.79 and the  $CC_{ML}$  was 0.53, while the  $ARI_{DWLS}$  was 0.80 and the  $CC_{DWLS}$  was 0.56. These results are remarkably similar to (but slightly worse than) the ones found in the main text (Section 4.3) for both versions of MMG-SEM. The main effects of the manipulated conditions on the cluster recovery were also largely the same. Specifically, higher within-group sample sizes  $N_g$ , larger  $\beta$  parameters, more response categories  $c$ , balanced cluster sizes, and fewer clusters  $K$  led to better performance for both versions of MMG-SEM. Notably, the main effects of the loading non-invariance size showed important differences compared to the main text.

Whereas the loading non-invariance size had minimal impact on cluster recovery when the non-invariances were modeled, it greatly impacts the results when the non-invariances are ignored. Specifically, the  $ARI_{ML}$  was 0.829 and 0.740, and the  $ARI_{DWLS}$  was 0.833 and 0.767 when  $\lambda$  non-invariance size was 0.2 and 0.4, respectively. Note that threshold non-invariance size showed almost no impact on the cluster recovery results.

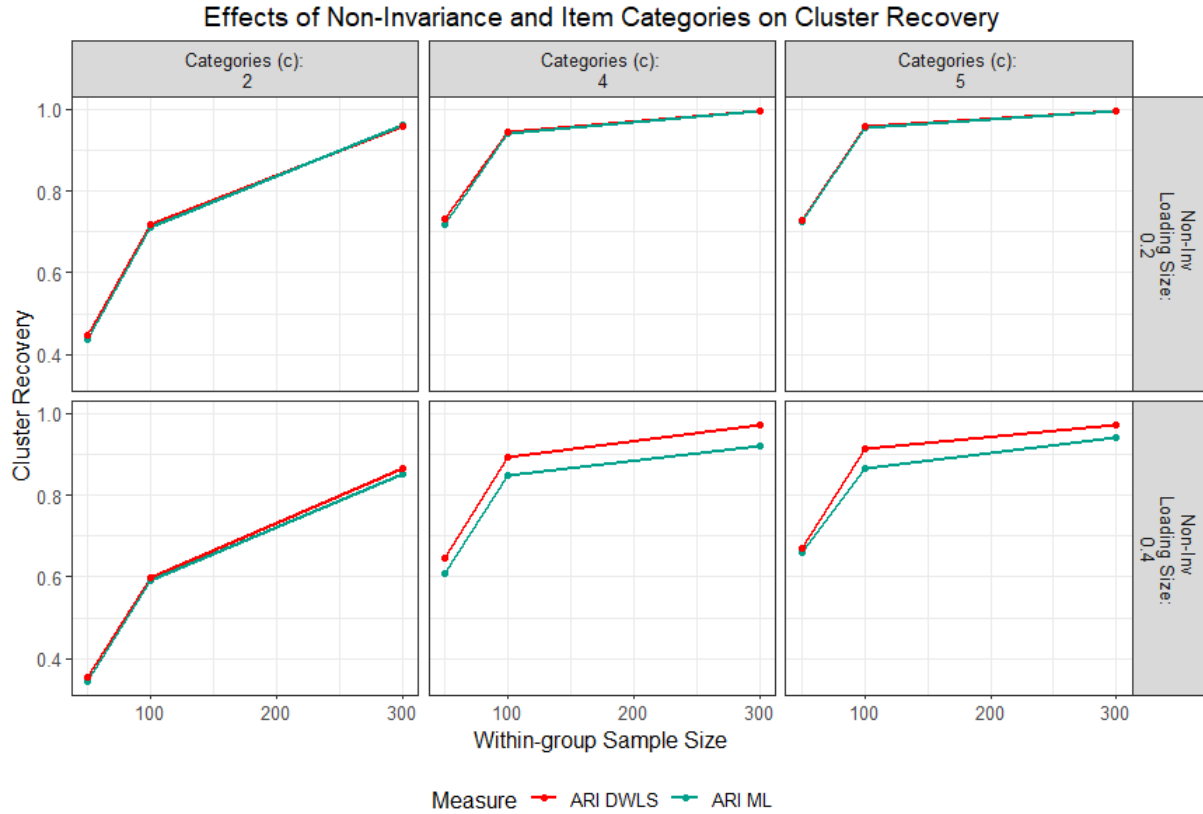
**Table B1**

*ARI and CC per level of each manipulated factor for both versions of MMG-SEM.*

Factor	Level	$ARI_{ML}$	$CC_{ML}$	$ARI_{DWLS}$	$CC_{DWLS}$
$\beta$ Size	0.2	0.634 (0.372)	0.325 (0.469)	0.644 (0.375)	0.349 (0.477)
	0.3	0.841 (0.249)	0.566 (0.496)	0.857 (0.240)	0.596 (0.491)
	0.4	0.894 (0.207)	0.687 (0.464)	0.911 (0.189)	0.731 (0.444)
$N_g$	50	0.576 (0.371)	0.242 (0.428)	0.589 (0.372)	0.266 (0.442)
	100	0.817 (0.254)	0.493 (0.500)	0.833 (0.248)	0.531 (0.499)
	300	0.945 (0.146)	0.802 (0.399)	0.960 (0.123)	0.836 (0.370)
$K$	2	0.872 (0.250)	0.671 (0.470)	0.875 (0.248)	0.687 (0.464)
	4	0.708 (0.334)	0.384 (0.486)	0.734 (0.333)	0.432 (0.495)
Cluster size	Bal	0.824 (0.287)	0.577 (0.494)	0.833 (0.285)	0.604 (0.489)
	Unbal	0.754 (0.321)	0.474 (0.499)	0.774 (0.316)	0.511 (0.500)
Non-inv Size $\lambda$	0.2	0.829 (0.291)	0.596 (0.491)	0.833 (0.287)	0.600 (0.490)
	0.4	0.740 (0.318)	0.439 (0.496)	0.767 (0.317)	0.505 (0.500)
Non-inv Size $\tau$	0.25	0.784 (0.310)	0.517 (0.500)	0.800 (0.307)	0.552 (0.497)
	0.50	0.794 (0.303)	0.533 (0.499)	0.800 (0.298)	0.553 (0.496)
Item categories $c$	2	0.656 (0.332)	0.291 (0.454)	0.664 (0.328)	0.291 (0.454)
	4	0.851 (0.269)	0.626 (0.484)	0.874 (0.254)	0.671 (0.470)
	5	0.877 (0.258)	0.691 (0.462)	0.890 (0.258)	0.746 (0.435)
Total	—	0.789 (0.307)	0.525 (0.499)	0.803 (0.303)	0.558 (0.497)

Note.  $\beta$  Size is the size of (difference in) regression parameters,  $N_g$  is the within-group sample size,  $K$  is the number of clusters, Non-inv Size  $\lambda$  is the size of the loadings' non-invariance, Non-inv Size  $\tau$  is the size of the thresholds' non-invariance, and  $c$  is the number of categories for each item.

To better understand the impact of the loading non-invariance size, we inspected its interaction with the most important manipulated factors in Figure B1. While it is clear that both versions of MMG-SEM decreased in performance in the case of larger non-invariances, the effect of ignoring the loadings non-invariance is more notorious when using the ML instead of the DWLS. Specifically, if the loadings non-invariance was 0.4,  $ARI_{ML}$  was consistently (slightly) lower than  $ARI_{DWLS}$  even when  $c > 2$ .

**Figure B1**

*Cluster Recovery in Function of Within-Group Sample Size, Loadings Non-invariance Size, and Item Categories.*

### 6.3 Regression Parameter Recovery

The recovery of the regression parameters was evaluated through  $RMSE_{\beta}$  (Equation 9). Table B2 shows the main effects of all manipulated factors on the  $RMSE_{\beta}$ . Overall, both  $RMSE_{\beta,ML}$  (0.051) and  $RMSE_{\beta,DWLS}$  (0.056) presented similar results. Furthermore, we found the same pattern of results as in the main text. Specifically, higher within-group sample sizes  $N_g$ , larger  $\beta$  parameters, more response categories  $c$ , balanced cluster sizes, and fewer clusters  $K$  led to lower RMSE values for both versions of MMG-SEM.

Remarkably, the trend observed in cluster recovery regarding the loadings non-invariance size is also evident in the regression parameter recovery. When ignoring the non-invariances, the performance of both versions of MMG-SEM decreased as the size of the loadings non-invariance increased. Specifically, the  $RMSE_{\beta,ML}$  was 0.041 and 0.064, and the  $RMSE_{\beta,DWLS}$  was 0.048 and 0.066 when the loadings non-invariance size was 0.2

and 0.4, respectively.

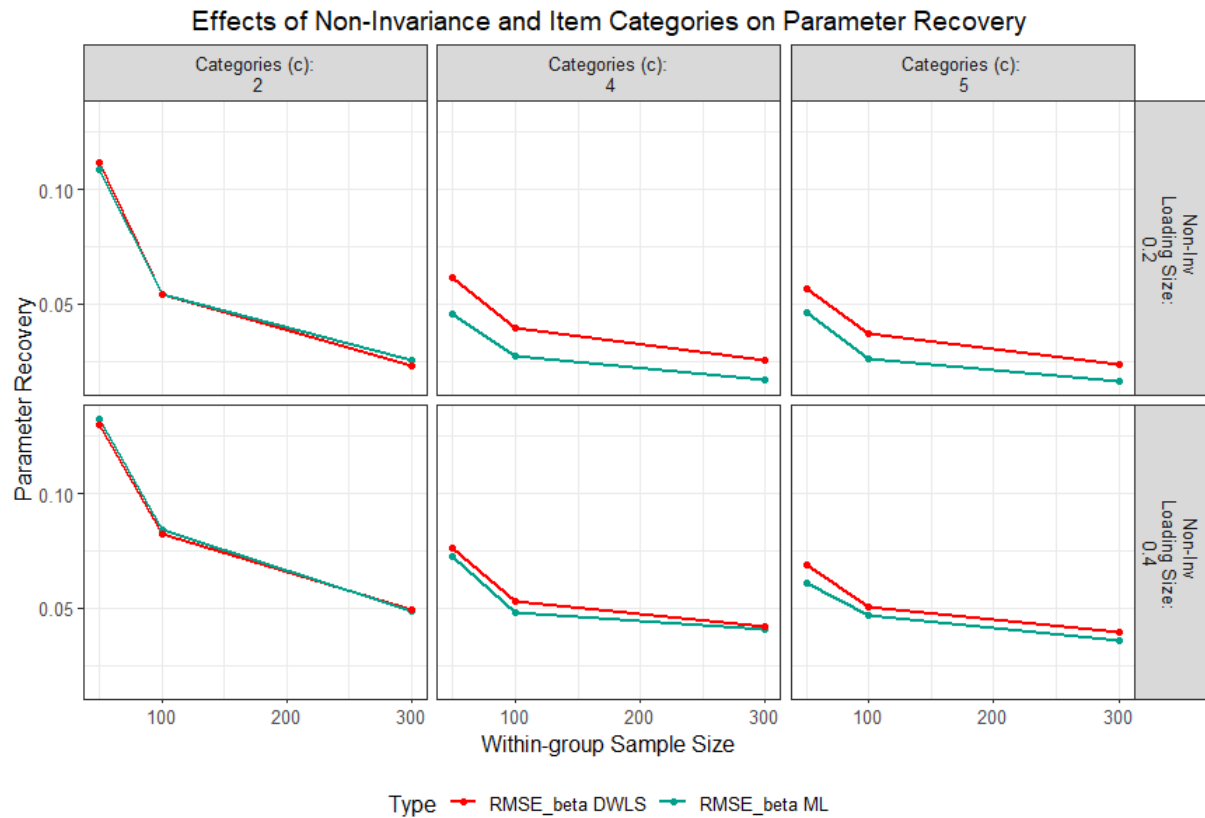
**Table B2**

RMSE $_{\beta}$  per level of each manipulated factor for both versions of MMG-SEM.

Factor	Level	RMSE $_{\beta,ML}$	RMSE $_{\beta,DWLS}$
$\beta$ Size	0.2	0.053 (0.054)	0.056 (0.054)
	0.3	0.049 (0.054)	0.054 (0.050)
	0.4	0.051 (0.054)	0.058 (0.051)
$N_g$	50	0.079 (0.070)	0.086 (0.068)
	100	0.047 (0.046)	0.053 (0.043)
	300	0.030 (0.028)	0.033 (0.024)
$K$	2	0.029 (0.026)	0.036 (0.027)
	4	0.072 (0.064)	0.075 (0.062)
Cluster size	Bal	0.039 (0.041)	0.045 (0.039)
	Unbal	0.063 (0.062)	0.067 (0.060)
Non-inv Size $\lambda$	0.2	0.041 (0.046)	0.048 (0.047)
	0.4	0.064 (0.060)	0.066 (0.056)
Non-inv Size $\tau$	0.25	0.052 (0.053)	0.056 (0.050)
	0.50	0.050 (0.055)	0.056 (0.054)
Item categories $c$	2	0.074 (0.071)	0.074 (0.072)
	4	0.040 (0.036)	0.048 (0.032)
	5	0.036 (0.033)	0.044 (0.031)
Total	—	0.051 (0.054)	0.056 (0.052)

Note.  $\beta$  Size is the size of the (difference in) regression parameters,  $N_g$  is the within-group sample size,  $K$  is the number of clusters, Non-inv Size  $\lambda$  is the size of the loadings' non-invariance, Non-inv Size  $\tau$  is the size of the thresholds' non-invariance, and  $c$  is the number of categories for each item.

The interaction of the loadings non-invariance size with the within-group sample size and the number of item categories is depicted in Figure B2. Similar to the results in the main text, the RMSE $_{\beta,ML}$  is consistently lower than RMSE $_{\beta,DWLS}$  in most conditions. Notably, both RMSE $_{\beta,ML}$  and RMSE $_{\beta,DWLS}$  performed considerably worse when the loadings non-invariance size increased, but the drop in performance was more notable for RMSE $_{\beta,ML}$  than for RMSE $_{\beta,DWLS}$ .

**Figure B2**

*Regression Parameter Recovery in Function of Within-Group Sample Size, Loadings Non-invariance Size, and Item Categories.*

## Appendix C

### Empirical Example - Treating Data as Continuous

This Appendix describes the results of the empirical example when the data was treated as continuous. Specifically, we used the standard ML estimation of MG-CFA in the first step of MMG-SEM. The dataset, research question, and the SEM model are described in detail in Section 5 and Figure 5.

Before applying MMG-SEM, we tested for measurement invariance (Vandenberg & Lance, 2000), for which the results are summarized in Table C1. Both the configural model (CFI = 0.971, RMSEA = 0.068) and the metric model (CFI = 0.966, RMSEA = 0.065) presented a satisfactory model fit, and the difference between them was deemed as acceptable ( $<0.01$ ) (Rutkowski & Svetina, 2014). Note that, because the data was treated as continuous, thresholds were not included in the model, making it impossible to test for their invariance.

**Table C1**

*Model fit resulting from the measurement invariance testing per level.*

Measurement Invariance level	CFI	RMSEA
Configural	0.971	0.068
Metric	0.966	0.065

After establishing metric invariance, we proceeded to analyze the data using MMG-SEM. As in the main text, the first step was model selection in terms of the number of clusters. Following the recommendations of Perez Alonso et al. (2025), we used  $AIC_3$  and the Convex Hull scree ratio. Unlike the analysis using MG-CCFA in the main text, which supported a 3-cluster solution, the model selection measures suggested a 2-cluster solution when treating the data as continuous (see the GitHub repository for more details).

The resulting clusters can be seen in Table C2. Although the number of clusters differed, the overall grouping of the countries was strikingly similar across the two approaches. When treating data as continuous, Cluster 1 primarily consisted of Northern and Eastern European countries, now also including South Korea, whereas Cluster 2 mainly comprised Western European countries. A few changes were observed: the Czech Republic, Austria, and Iceland were grouped with the Eastern European cluster, and the

countries in the Americas (i.e., Mexico and Colombia) were grouped with the Western European countries.

**Table C2**

*Clustering of the countries based on the regression parameter between parental emotional support and self-efficacy (2-cluster model).*

Cluster	Countries
Cluster 1	South Korea, Austria, Czech Republic, Estonia, Finland, Greece, Hungary, Iceland, Lithuania, Luxembourg, Slovenia, Sweden, Turkey, United States of America
Cluster 2	Switzerland, Chile, Colombia, Germany, Spain, France, United Kingdom, Ireland, Japan, Latvia, Netherlands, Poland, Portugal, Slovakia

In terms of the regression parameters (Table C3), the estimates were slightly smaller than those in the main text. However, the overall trend was consistent. Both clusters presented positive effects, with the Eastern European cluster (i.e., Cluster 1) showing a larger coefficient compared to the Western European cluster.

**Table C3**

*Regression parameters per cluster for the relation between parental emotional support and self-efficacy.*

	Cluster 1	Cluster 2
Regression parameter	0.218	0.166

In summary, the results of MMG-SEM using ML estimation in the first step were broadly similar to those obtained using DWLS. The main difference was the model selection step, which identified a different number of clusters. This raises relevant questions about the impact of treating ordinal data as continuous on model selection measures. Such an issue requires further research in the future.