Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis

Kim De Roover

Tilburg University

Author Notes:

The research leading to the results reported in this paper was funded by the Netherlands Organization for Scientific Research (NWO) [Veni grant 451-16-004]. Correspondence concerning this paper should be addressed to Kim De Roover, Tilburg School of Social and Behavioral Sciences, Department of Methodology and Statistics, PO box 90153 5000 LE Tilburg, The Netherlands. E-mail: K.DeRoover@uvt.nl.

Abstract

Comparisons of latent constructs across groups are ubiquitous in behavioral research and, nowadays, often numerous groups are involved. Measurement invariance of the constructs across the groups is imperative for valid comparisons and can be tested by multigroup factor analysis. Metric invariance (invariant factor loadings) often holds, whereas scalar invariance (invariant intercepts) is rarely supported across many groups. Scalar invariance is a prerequisite for comparing latent means, however. One may inspect group-specific intercepts to pinpoint non-invariances, but this is a daunting task in case of many groups. This paper presents mixture multigroup factor analysis (MMG-FA) for clustering groups based on their intercepts. Clusters of groups with scalar invariance are obtained by imposing cluster-specific intercepts and invariant loadings whereas unique variances, factor means and factor (co)variances can differ between groups. Thus, MMG-FA ties down the number of intercepts to inspect and generates clusters of groups wherein latent means can be validly compared.

Keywords: Measurement invariance, multigroup factor analysis, strong invariance, scalar invariance, mixture modeling.

1. Introduction

Behavioral researchers are commonly interested in latent constructs such as personality traits or attitudes, measured by questionnaire items (or other indicators). Often, the goal is to make between-group comparisons with regard to the constructs, for instance, to assess cross-cultural differences in human values (Inglehart, Basanez, Basanez, & Moreno, 1998). Before doing so, testing for measurement invariance (MI) across the groups is imperative (Meredith, 1993). Specifically, a ‘measurement model’ (MM) indicates how the constructs are measured by the

questionnaire items and measurement invariance holds when the MM is equal across groups. If so, the constructs are measured in exactly the same way across groups and are thus validly comparable. If not, between-group comparisons are invalidated (Chen, 2008; Greiff, & Scherer, 2018).

For continuous items, or items that are treated as such, the MM is traditionally evaluated with factor analysis (Lawley & Maxwell, 1962), where the extent to which an item measures a construct or 'factor' is indicated by a 'factor loading'. Confirmatory factor analysis (CFA) is used to impose a priori assumptions about which items are measuring which factors (by allowing a pre-specified subset of loadings to differ from zero) and evaluate the fit of this MM for a given data set. When such prior knowledge is lacking or one wants to explore the measured constructs without imposing zero loadings, exploratory factor analysis (EFA) is used. Regardless of the MM being assessed by CFA or EFA, measurement invariance pertains to the equality of certain parameters of the factor model across groups. Measurement invariance can be tested by multigroup factor analysis (MG-FA; Dolan, Oort, Stoel, & Wicherts, 2009; Jöreskog, 1971; Sörbom, 1974). Specifically, invariance of MM parameters holds when the decrease in model fit caused by imposing their equality across groups is non-significant (Cheung & Rensvold, 2002). In multigroup CFA, one first inspects the model fit for the baseline or 'configural invariance' model, to evaluate whether the number of factors and the imposed pattern of zero loadings holds across the groups (Meredith, 1993). In multigroup EFA, no specific zero loadings are imposed. In both approaches, the tenability of 'weak' or 'metric invariance' is evaluated by imposing equal factor loadings across groups. When metric invariance holds, latent structures (e.g., how values affect immigration attitudes; Beierlein, Kuntz, & Davidov, 2016) are comparable across groups. Next, 'strong' or 'scalar invariance' is tested by also restricting the item intercepts to be equal across

groups. Scalar invariance is a prerequisite for comparing latent means across groups. Finally, 'strict invariance' pertains to the invariance of the items' residual or 'unique' variances.

When a certain level of measurement invariance is rejected across groups, one may attempt to pinpoint sources of non-invariance – i.e., problematic parameters and/or groups – by comparing the group-specific factor models. However, the number of pairwise comparisons grows exponentially with an increasing number of groups and, nowadays, many studies involve a lot of groups (Kim, Cao, Wang, & Nguyen, 2017; Rutkowski & Svetina, 2014); for example, cross-national surveys such as the World Values Survey. The many comparisons make it hard to unravel invariant from non-invariant parameters and for which groups they apply (Byrne & van de Vijver, 2010) and elevates the chances of falsely detecting non-invariance (Rutkowski & Svetina, 2014).

The large number of groups makes it unlikely that all groups have the same MM, but it is equally unlikely that each group has its own MM. Realistically, some groups may have (near-)identical measurement parameters and, therefore, a few latent clusters of groups may emerge (Byrne & van de Vijver, 2010). To capture these latent clusters, De Roover, Vermunt and Ceulemans (2020) recently presented 'mixture multigroup factor analysis' (MMG-FA), which is an extension of MG-FA that performs a mixture clustering (McLachlan & Peel, 2000) of the groups based on (a specific subset of) the MM parameters. As described above, different levels of measurement (non-)invariance have different implications in terms of which comparisons are (in)valid (Meredith, 1993) and, therefore, MMG-FA clusters groups based on their MM parameters in a level-specific way. In fact, De Roover, Vermunt and Ceulemans (2020) focused on a variant of MMG-FA that clusters the groups on their factor loadings, to find clusters of groups for which metric invariance holds. For empirical data on many groups, scalar invariance is rarely supported, however, whereas metric invariance is established frequently. In cross-cultural research, for

instance, scalar non-invariance is more often a threat to valid cross-cultural comparisons than metric non-invariance (Boer, Hanke, & He, 2018; Davidov, Dülmer, Schlüter, Schmidt, & Meuleman, 2012). Therefore, this paper presents a variant of MMG-FA that clusters the groups on their intercepts specifically, whereas factor loadings are assumed to be invariant across groups.

By identifying clusters of groups with the same intercepts, MMG-FA (1) ties down the number of intercepts to compare, making it easier to identify non-invariant items, (2) indicates for which groups scalar invariance holds, allowing for valid comparisons of the latent means within each cluster, and (3) indicates potentially interesting between-group differences, for instance, cross-cultural differences in the functioning of an item. Obviously, the user somehow needs to determine the appropriate number of clusters for a given data set. A solution to this model selection problem was already discussed by De Roover et al. (2020).

Note that clustering the groups based on all MM parameters at once (i.e., factor loadings, item intercepts and residual or 'unique' variances) would imply the stringent assumption that the same clustering is underlying all MM parameters, whereas some parameter differences may be explained by another clustering – possibly with more clusters. When this assumption does not hold, such a clustering may fail to capture any of the MM differences properly or would need many clusters to do so. For the same reason, MMG-FA also sets aside differences in 'structural' parameters, i.e., factor (co)variances and factor means, as they are irrelevant to the MI question. This is exactly what makes MMG-FA unique to other approaches for evaluating measurement (non-)invariance across many groups (for an overview, see Kim et al., 2017), i.e., it is the only method that clusters the groups exclusively on specific subsets of the MM parameters.

The remainder of this paper is organized as follows: Section 2 discusses MG-FA and its extension into MMG-FA for intercept non-invariance. Section 3 describes a simulation study to

evaluate the performance of MMG-FA, including model selection. Section 4 illustrates its empirical value by scrutinizing intercept non-invariance for human values data from the European Social Survey. Section 5 concludes with points of discussion and directions for future research.

2. Method

2.1. Multigroup factor analysis

Multigroup factor analysis (MG-FA) operates on data from multiple groups, for example, age groups, religions, patient groups, or countries. Formally, the groups are indicated by $g = 1, \ldots, G$ and the subjects within the groups by $n_g = 1, \ldots, N_g$. The item scores for subject n_g on the J items are denoted by the J-dimensional vector $\mathbf{x}_{n_g}$ and, per group g, they are gathered into an $N_g \times J$ matrix $\mathbf{X}_g$. The factor model for the scores in $\mathbf{x}_{n_g}$ is written as (Lawley & Maxwell, 1962):

$$\mathbf{x}_{n_g} = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\eta}_{n_g} + \boldsymbol{\varepsilon}_{n_g} \tag{1}$$

where $\boldsymbol{\tau}_g$ indicates a J-dimensional group-specific intercept vector, $\boldsymbol{\Lambda}_g$ denotes a $J \times Q$ matrix of group-specific factor loadings, $\boldsymbol{\eta}_{n_g}$ is a Q-dimensional vector of the subject's scores on the Q factors and $\boldsymbol{\varepsilon}_{n_g}$ is a J-dimensional vector of residuals. The factor loadings indicate how the factors are measured by the items. Per group, the factor scores are assumed to be identically and independently distributed (i.i.d.) as $MVN\left(\boldsymbol{\alpha}_g, \boldsymbol{\Phi}_g\right)$, independently of $\boldsymbol{\varepsilon}_{n_g}$, which are i.i.d. as $MVN\left(\mathbf{0}, \boldsymbol{\Psi}_g\right)$. The factor means of group g are denoted by $\boldsymbol{\alpha}_g$, whereas $\boldsymbol{\Phi}_g$ pertains to the group-specific factor (co)variances and $\boldsymbol{\Psi}_g$ to a diagonal matrix containing the group's unique variances. The model-implied covariance matrix for group g is $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$. Estimating Equation 1 for each group corresponds to the baseline model for MI testing. To partially identify the model, the factor means $\boldsymbol{\alpha}_g$ are fixed to zero and the factor variances on the diagonal of $\boldsymbol{\Phi}_g$ to one per group g.

Note that multigroup CFA (MG-CFA) tests invariance for an assumed MM by imposing specific zero loadings on $\mathbf{\Lambda}_g$. The tenability of this pattern of zero loadings across the groups is called 'configural invariance'. Note that, when configural invariance fails, resorting to EFA is often a better strategy (Gerbing & Hamilton, 1996) than elaborately respecifying CFA models, to avoid capitalization on chance (Browne, 2001; MacCallum, Roznowski, & Necowitz, 1992). Moreover, fixed zero loadings are often too restrictive (Asparouhov & Muthén, 2009; McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Muthén & Asparouhov, 2012) and model misspecifications can severely bias other parameter estimates (Anderson & Gerbing, 1982; Bollen, Kirby, Curran, Paxton, & Chen, 2007). As an alternative, multigroup EFA (MG-EFA) does not impose zero loadings, but includes rotational freedom of the factors (De Roover & Vermunt, 2019; Dolan, Oort, Stoel, & Wicherts, 2009; Hessen, Dolan, & Wicherts, 2006).

To test for weak or metric invariance, the fit of the model with invariant loadings, i.e., $\mathbf{\Lambda}_g = \mathbf{\Lambda}$ for g = 1, …, G, is compared to that of the baseline model. For the metric invariance model, factor variances of one are imposed for one group or for the mean factor variances across groups; i.e., $diag\left(\frac{1}{N}\sum_{g=1}^{G} N_g \mathbf{\Phi}_g\right) = \mathbf{1}$ where $\mathbf{1}$ is a vector of ones and $N = \sum_{g=1}^{G} N_g$. To evaluate the comparability of latent means, strong or scalar invariance is tested by also restricting the intercepts to be invariant across groups, i.e., $\boldsymbol{\tau}_g = \boldsymbol{\tau}$ for g = 1, …, G, while freely estimating factor means $\boldsymbol{\alpha}_g$ for all groups but one (Dolan et al., 2009; Meredith, 1993). Specifically, for model identification, $\boldsymbol{\alpha}_g = \mathbf{0}$ for one group or $\frac{1}{N}\sum_{g=1}^{G} N_g \boldsymbol{\alpha}_g = \mathbf{0}$ across groups. This paper focuses on the situation where metric invariance holds but scalar invariance fails. Therefore, the next step of evaluating strict invariance, i.e., $\mathbf{\Psi}_g = \mathbf{\Psi}$ for g = 1, …, G, is beyond the scope of this paper.

To evaluate whether the fit of MG-FA is significantly worse with invariant intercepts, comparing fit indices such as the CFI and RMSEA – to evaluate 'practical' significance – is recommended since χ^2- difference tests for nested models are strongly affected by sample size. Specifically, non-invariance of intercepts is indicated when the decrease in CFI (ΔCFI) exceeds .01 and the increase in RMSEA (ΔRMSEA) is larger than .01 when imposing invariant intercepts (Chen, 2007; Cheung & Rensvold, 2002). When scalar invariance indeed fails, one can return to the metric invariance model, i.e., with invariant loadings and group-specific intercepts, and compare the intercepts among groups to locate non-invariances. This quickly becomes cumbersome when more than a few groups are involved, however. For instance, comparing intercepts for five groups implies only 10 pairwise comparisons, but 10 groups already require 45 comparisons and 25 groups result in no less than 300 comparisons. The next subsection presents mixture multigroup factor analysis for tying down the number of comparisons needed to identify intercept non-invariances.

2.2. Mixture multigroup factor analysis

2.2.1. Model specification

Mixture multigroup factor analysis (MMG-FA) aims to gather groups into clusters according to the equivalence of their MM parameters such that a specific level of measurement invariance holds within each cluster. This paper focuses on scalar (non-)invariance and, thus, on finding clusters of groups with invariant intercepts. The observations $\mathbf{x}_{n_g}$ are assumed to be sampled from a mixture of *K* multivariate normal distributions where all observations of a group are sampled from the same distribution. Thus, the mixture clustering operates at the group level. The *K* mixture components are henceforth called 'clusters'. Formally, the MMG-FA model for group *g* is written as:

$$f\left(\mathbf{X}_g;\boldsymbol{\theta}\right)=\sum_{k=1}^{K}\pi_k f_{gk}\left(\mathbf{X}_g;\boldsymbol{\theta}_{gk}\right)=\sum_{k=1}^{K}\pi_k\prod_{n_g=1}^{N_g}MVN(\mathbf{x}_{n_g};\boldsymbol{\mu}_{gk},\boldsymbol{\Sigma}_g)$$
$$with\quad \boldsymbol{\Sigma}_g=\boldsymbol{\Lambda}\boldsymbol{\Phi}_g\boldsymbol{\Lambda}'+\boldsymbol{\Psi}_g\quad and\quad \boldsymbol{\mu}_{gk}=\boldsymbol{\tau}_k+\boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk} \tag{2}$$

where *f* is the total population density function, and $\boldsymbol{\theta}$ refers to the total set of parameters. Similarly, f_{gk} refers to the *k*th cluster-specific density function for group *g* and $\boldsymbol{\theta}_{gk}$ refers to the corresponding set of parameters. The mixing proportions (i.e., prior probabilities of belonging to each of the clusters) are indicated by π_k, with $\sum_{k=1}^{K}\pi_k=1$. It is important to note that the covariance matrices are group-specific whereas the means are both group- and cluster-specific. Specifically, the covariance matrices are modeled by means of *invariant* factor loadings $\boldsymbol{\Lambda}$, *group-specific* factor (co)variances $\boldsymbol{\Phi}_g$, and *group-specific* unique variances $\boldsymbol{\Psi}_g$. The means are modeled by *cluster-specific* intercepts $\boldsymbol{\tau}_k$ and *group- and cluster-specific* factor means $\boldsymbol{\alpha}_{gk}$. This combination of invariant, group- and cluster-specific parameters assures that the clustering of groups is driven exclusively by intercept differences. It is notable that $\boldsymbol{\alpha}_{gk}$ is not only group-specific but also varies across clusters. Assuming a group's factor means to be the same in all clusters would be too restrictive, because they depend on the cluster-specific intercepts. As shown in Appendix A, it is possible to estimate the factor means for every group-and-cluster combination despite the fact that the mixture model assumes that each group belongs to only one cluster. The resulting $\boldsymbol{\alpha}_{gk}$ contains the factor means conditional on group *g* belonging to cluster *k*. For each group *g*, the factor means for the clusters the group does *not* belong to are nuisance parameters.

Thus, in MMG-FA, the factor model is conditional on the cluster membership of group *g*, indicated by z_{gk}, as follows:

$$\left[\mathbf{x}_{n_g}\mid z_{gk}=1\right]=\boldsymbol{\tau}_k+\boldsymbol{\Lambda}\boldsymbol{\eta}_{n_g k}+\boldsymbol{\varepsilon}_{n_g} \tag{3}$$

where $\boldsymbol{\eta}_{n_g k} \sim MVN\left(\boldsymbol{\alpha}_{gk}, \boldsymbol{\Phi}_g\right)$ and $\boldsymbol{\varepsilon}_{n_g} \sim MVN\left(\mathbf{0}, \boldsymbol{\Psi}_g\right)$. Note that, whereas De Roover et al. (2020) used EFA within clusters to scrutinize loading non-invariances, it is possible to use either CFA or EFA for modeling the invariant loadings $\boldsymbol{\Lambda}$ in the MMG-FA variant proposed in this paper. Indeed, when assumed zero loadings do not hold or are overly restrictive (see Section 2.1), MMG-EFA is a good alternative and, in that case, the factors can be rotated to facilitate interpretation. For each factor, the scale is set by fixing the mean of the group-specific factor variances to one across all groups. Since it is unknown beforehand which groups belong to a cluster, we refrain from imposing factor means of zero for one group per cluster. Instead, we restrict the factor means per cluster as follows: $\frac{1}{N_k}\sum_{g=1}^{G} N_g \hat{z}_{gk} \boldsymbol{\alpha}_{gk} = \mathbf{0}$ where $N_k = \sum_{g=1}^{G} N_g \hat{z}_{gk}$ and $\hat{z}_{gk}$ indicates the estimated cluster memberships or posterior classification probabilities (Appendix A). Note that the factor means $\boldsymbol{\alpha}_{gk}$ can be compared among groups assigned to the same cluster, thus providing researchers with the latent mean comparisons they were looking for – albeit per cluster of groups.

Note that the existing method that is most similar to MMG-FA (as specified above) is multilevel factor mixture modeling (ML-FMM; Kim et al., 2017). Unlike MMG-FA, it clusters the groups based on measurement and structural parameters at the same time (and only allows to use CFA). When metric invariance holds and the aim is to trace intercept non-invariances, ML-FMM is specified such that the factor loadings are invariant and the clustering is driven by intercept differences as well as differences in unique variances, factor means, and factor (co)variances. This implies that unique variances, factor means, and factor (co)variances are assumed to be the same for all groups within a cluster, which is too restrictive when looking for clusters of groups wherein scalar invariance holds. ML-FMM is thus less focused on intercept differences than MMG-FA.

2.2.2. Model estimation

The unknown parameters $\boldsymbol{\theta}$ of the MMG-FA model are estimated by means of maximum likelihood (ML) estimation. This involves maximizing the logarithm of the likelihood of the data:

$$\begin{aligned}\log L &= \log\left(\prod_{g=1}^{G}\sum_{k=1}^{K}\pi_k\prod_{n_g=1}^{N_g}\frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Sigma}_g\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_{n_g}-\boldsymbol{\mu}_{gk}\right)'\boldsymbol{\Sigma}_g^{-1}\left(\mathbf{x}_{n_g}-\boldsymbol{\mu}_{gk}\right)\right)\right)\\ &= \sum_{g=1}^{G}\log\left(\sum_{k=1}^{K}\pi_k\prod_{n_g=1}^{N_g}\frac{1}{(2\pi)^{J/2}\left|\boldsymbol{\Sigma}_g\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_{n_g}-\boldsymbol{\mu}_{gk}\right)'\boldsymbol{\Sigma}_g^{-1}\left(\mathbf{x}_{n_g}-\boldsymbol{\mu}_{gk}\right)\right)\right),\end{aligned} \tag{4}$$

where $\boldsymbol{\mu}_{gk}$ and $\boldsymbol{\Sigma}_g$ are decomposed as specified in Equation 2. Note that obtaining the parameter estimates $\hat{\boldsymbol{\theta}}$ by means of Newton-Raphson, Fisher scoring or Quasi-Newton optimization methods – i.e., methods that are used in commercial software such as Latent Gold (Vermunt & Magidson, 2013, 2016) and Mplus (Muthén & Muthén, 2005) – is very slow due to the large number of parameters and sensitive to starting values. To find the parameter estimates in a time-efficient and stable manner, we developed an alternating expectation-conditional maximization (AECM) algorithm (Appendix A), which has been implemented in Matlab R2019b, R (see https://github.com/KimDeRoover/MixtureMG_FA) and Latent Gold 6.0 (Appendix B). Because the algorithm may end up in a local maximum, it applies a multistart procedure to increase the probability of obtaining the global maximum.

2.2.3. Model selection

In this paper, we focus on the case where the number of factors is assumed to be known. Thus, the model selection problem is confined to selecting the most appropriate number of clusters *K* for a given data set. For related mixture models, minimizing the Bayesian Information Criterion (BIC; Schwarz 1978) is often recommended (Nylund, Asparouhov, & Muthén, 2007; Tay, Diener, Drasgow, & Vermunt, 2011; Tein, Coxe, & Cham, 2013). The BIC balances the log *L* and model complexity by penalizing a model with more free parameters *fp* and larger sample size as follows:

$$\text{BIC} = -2\log L + fp\log(N) \quad (5)$$

For MMG-FA, *fp* is equal to the sum of the number of mixing proportions (minus one restriction), intercepts, factor loadings (minus *R* restrictions), factor variances (minus *Q* restrictions), factor covariances, factor means (minus *KQ* restrictions, excluding nuisance parameters), and unique variances: $fp = K - 1 + KJ + (JQ - R) + (G-1)Q + GQ(Q-1)/2 + (G-K)Q + GJ$. For CFA, '*R*' equals the number of imposed zero loadings and, for EFA, it equals $Q(Q-1)$ (i.e., the correction for rotational freedom). Several authors (Kim, Joo, Lee, Wang, & Stark, 2016; Lukočienė, Varriale, & Vermunt, 2010) suggested that, for group-level clusters, one should use the number of groups *G* – instead of *N* – for the sample size in the BIC computation. For small sample sizes and low cluster separation in multilevel mixture modeling, the Akaike Information Criterion (AIC; Akaike, 1973) was found to outperform BIC (Kim et al., 2017; Lukočienė & Vermunt, 2010). For growth mixture models, Bauer (2007) and McNeish and Harring (2017) indicated that in less ideal but empirically realistic conditions (e.g., non-normality), BIC and AIC tend to overselect clusters.

On top of that, De Roover et al. (2020) showed that the Convex Hull procedure (CHull) is a valuable alternative to BIC and AIC for selecting the number of clusters in MMG-FA. The CHull (Ceulemans & Van Mechelen, 2005; Ceulemans & Kiers, 2006) is a generalization of the scree test (Cattell, 1966) based on a plot of the log *L* versus *fp* of the obtained solutions. Specifically, for each solution on the convex hull of the plot, the following scree ratio is computed and the solution with the maximal scree ratio is selected: $\frac{\log L_K - \log L_{K+1}}{fp_K - fp_{K+1}} \Big/ \frac{\log L_{K-1} - \log L_K}{fp_{K-1} - fp_K}$. Note that, like a scree test, CHull cannot select the least complex model and thus selects at least two clusters. But when focusing on cases where scalar invariance was rejected, and intercept differences are thus expected to be present, this is not a problem. Furthermore, visual inspection of the CHull plot

may lead to the conclusion that no clear elbow is present and thus that an underlying clustering is unlikely. In the next section, these methods are compared in terms of their performance with regard to selecting the correct number of clusters underlying the between-group intercept differences.

3. Simulation Study

3.1. Problem

On the one hand, the goal of the simulation study is to evaluate the performance of MMG-FA in terms of recovering the clustering of the groups and the parameters when the correct number of clusters is known. On the other hand, it is evaluated to what extent the model selection procedures described in Section 2.2.3 select the correct number of clusters for MMG-FA. We manipulated seven factors that were expected to affect the cluster separation and/or the stability of parameter estimates, and thus the performance of MMG-FA and its model selection: (1) the number of groups, (2) the group sizes, (3) the number of clusters, (4) the cluster sizes, (5) the number of factors, and the size (6) and number (7) of intercept differences.

Specifically, in terms of their effect, we hypothesize the following: The number of groups (1) on the one hand determines how many groups end up within each cluster. Because more groups within a cluster implies more information on the cluster-specific intercepts (i.e., a higher within-cluster sample size), the performance may improve with a higher number of groups. On the other hand, a higher number of groups implies more cluster memberships (posterior classification probabilities) to be estimated, which may be more intricate. More observations per group (2) increase the within-cluster sample size and thus the performance. It also implies a higher cluster separation because more information is available for estimating each of the cluster memberships (Lukočienė, Varriale, & Vermunt, 2010). A higher number of clusters (3) lowers the within-cluster sample size (for a given number of groups) and is thus expected to lower the performance. It also

increases the number of cluster memberships to be determined for each group, making their recovery more intricate. The cluster sizes (4), corresponding to the mixing proportions, pertain to the groups being equally or unequally divided across the clusters. In the unequal case, larger cluster(s) compete with smaller cluster(s) and the smaller ones may be harder to recover both in terms of cluster memberships and intercepts. A higher number of factors (5) implies a lower factor overdetermination, given the same number of variables. In that case, more factor means need to be recovered as well, which may be more difficult. Finally, the size (6) and number (7) of intercept differences greatly determine the extent to which the cluster-specific intercepts differ from one another (i.e., cluster separation) and thus affects the recovery of the cluster memberships.

3.2. Design

These factors were systematically varied in a complete factorial design:

1. the *number of groups G* at 2 levels: 12, 60;
2. the *group sizes* N_g (i.e., number of subjects per group) at 4 levels: 50, 100, 300, 500;
3. the *number of clusters K* at 2 levels: 2, 4;
4. the *cluster sizes* at 2 levels: equal, unequal;
5. the *number of factors Q* at 2 levels: 2, 4;
6. the *size of intercept differences* at 2 levels: .60, .30;
7. the *number of intercept differences* at 2 levels: 8 or 2 per pair of clusters.

The number of variables *J* was fixed at 20 and the invariant factor loading matrix $\mathbf{\Lambda}$ was manipulated as follows:

$$\mathbf{\Lambda}' = \begin{bmatrix} \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} \end{bmatrix}$$

or

$$\mathbf{\Lambda}' = \begin{bmatrix} \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} & \sqrt{.6} \end{bmatrix}$$

for two and four factors, respectively. The non-zero loadings are equal to $\sqrt{.60}$ in order to obtain total variances that vary around one, since the unique variances vary around .40 (see below).

The numbers of groups of 12 and 60 correspond to the range of group numbers that are generally encountered in large-scale surveys (Rutkowski & Svetina, 2014). To obtain equal cluster sizes, the groups are equally divided across clusters, i.e., each cluster contains 50% of the groups when $K = 2$, and 25% when $K = 4$. In the unequal cluster size conditions, the groups are divided over the clusters such that one cluster contains 75% of the groups, whereas the rest is equally divided over the remaining clusters. Thus, in case of two clusters, 75% of the groups are in one cluster and 25% in the other one. When $K = 4$, the three smaller clusters contain 8.33% of the groups each, which implies that they are singleton clusters (i.e., including only one group) in case of 12 groups. In case of 60 groups, they hold five groups each. The group sizes range from 50 to 500, which correspond to very small and large sample sizes for factor analysis, respectively (Fabrigar, MacCallum, Wegener, & Strahan, 1999; MacCallum, Widaman, Zhang, & Hong, 1999). In the current study, MMG-FA focuses on differences in intercepts and the factor loadings are invariant across groups. Therefore, the accuracy of the cluster-specific intercept estimates is determined by the sample size of a cluster of groups, rather than of a single group, whereas the accuracy of the factor loadings depends on the total sample size across all groups. The precision of the group- and cluster-specific factor means still depends on the group sizes, however.

The cluster-specific intercepts $\boldsymbol{\tau}_k$ are zero for the invariant items and equal to .60 or .30 for the non-invariant items. An intercept difference of .60 is considered to be a large non-invariance (Meade, Johnson, & Braddy, 2008; Stark, Chernyshenko, & Drasgow, 2006; Woods, 2009) that, when ignored, can lead to incorrect statistical inferences and biased parameter estimates (Hancock, Lawrence, & Nevitt, 2000; Kim & Willson, 2014). To obtain two intercept differences

for each pair of clusters, the intercept was increased to .60 or .30 for one item per cluster. Similarly, altering four intercepts per cluster led to eight differences per cluster pair. The set of non-invariant items differs across the clusters: e.g., in case of eight differences per cluster pair and four clusters, the non-invariant items are [1, 6, 11, 16] for cluster 1, [2, 7, 12, 17] for cluster 2, [3, 8, 13, 18] for cluster 3 and [4, 9, 14, 19] for cluster 4. The group- and cluster-specific factor means $\boldsymbol{\alpha}_{gk}$ are randomly sampled from a uniform distribution between −.50 and .50, i.e., from $U(-.50,.50)$. To make sure that the clustering is driven only by the manipulated intercept differences, the factor means are 'centered' for each cluster *k* such that $\frac{1}{N_k}\sum_{g=1}^{G} N_g z_{gk} \boldsymbol{\alpha}_{gk} = \mathbf{0}$. Note that, for a given pair of groups, the factor mean differences range from 0 to 1.00, whereas a factor mean difference of .50 is considered to be substantial (Kim et al., 2017). The group-specific factor correlations are sampled[1] from $U(-.50,.50)$ and factor variances from $U(.50,1.50)$. Whenever a resulting $\boldsymbol{\Phi}_g$ is not positive definite, the sampling is repeated. Group-specific unique variances (i.e., the diagonal of $\boldsymbol{\Psi}_g$) are sampled from $U(.20,.60)$. Finally, factor scores are sampled from $MVN(\boldsymbol{\alpha}_{gk}, \boldsymbol{\Phi}_g)$ and residuals from $MVN(\mathbf{0}, \boldsymbol{\Psi}_g)$ and the simulated data are created according to Equation 3.

According to this procedure, 50 data sets were generated per cell of the design, using Matlab R2019b. Thus, 2 (number of groups) × 4 (group sizes) × 2 (number of clusters) × 2 (cluster sizes) × 2 (number of factors) × 2 (size of intercept differences) × 2 (number of intercept differences) × 50 (replications) = 12,800 data sets were generated. The data were analyzed by the AECM algorithm for MMG-FA detailed in Appendix A, using CFA with the correct number of factors *Q*, the correct pattern of zero loadings and using 10 starts (i.e., starting from 10 random

[1] Note that the random sampling of factor means and factor (co)variances is repeated for each simulated data set in order to promote generalization of the conclusions across different values for these parameters.

partitions of the groups). On the one hand, the correct number of clusters *K* was specified to evaluate the performance of the MMG-CFA algorithm itself. On the other hand, for the first 25 replications of each cell of the design (i.e., for 6,400 data sets), we performed MMG-CFA analyses with one to six clusters to evaluate the performance of the model selection procedures described in Section 2.2.3 and the goodness-of-recovery for MMG-CFA models with an incorrect number of clusters. To check the performance of the EFA-based algorithm, each data set was also analyzed with MMG-EFA with the correct *Q* and *K* and 10 starts. To compare the performance of MMG-CFA to that of ML-FMM, we applied ML-FMM with invariant loadings and the correct *K* and *Q* to all data sets (with 10 starts). All analyses converged within 1,000 iterations. The mean CPU time for MMG-FA with the correct *K* was 48 seconds, both for the CFA- and EFA-based algorithms, on an i5 processor with a clock frequency of 2.4 GHz and 8GB RAM. For the model selection, estimating the models with one to six clusters took about 5 minutes on average.

3.3. Results

Detailed results for the MMG-CFA analyses are discussed in Sections 3.3.1 to 3.3.5, whereas Section 3.3.6 gives an overview of the performance of MMG-EFA. The model selection results are discussed in Section 3.3.7 and the goodness-of-recovery for an incorrect number of clusters in Section 3.3.8.

3.3.1. Sensitivity to local maxima

To evaluate the frequency of local maximum solutions, we should compare the log *L* value of the best solution obtained by the multistart procedure (i.e., starting from 10 random partitions; see Appendix A2) with the global ML solution for each simulated data set. Because of sampling fluctuations, the global maximum is unknown, however. Therefore, we used a ‘proxy’ of the global ML solution; i.e., the solution that is obtained when the algorithm starts from the true clustering

of the groups. The best solution from the multistart procedure is then considered to be a local maximum when its log *L* value is smaller than the one from the proxy. To exclude mere calculation precision differences, we only considered such differences with an absolute value higher than .0001 to indicate a local maximum. By this definition, 1,296 (10.1%) local maxima were detected over all 12,800 simulated data sets. Not surprisingly, most of these occur in the more difficult conditions; e.g., 1,264 of the 1,296 local maxima are found in the conditions with four unequal clusters (i.e., with three very small clusters). Note that, for 919 out of these 1,296 data sets, re-running the analysis with 50 starts (i.e., starting from 50 random partitions) was sufficient to avoid the local maximum, reducing the percentage of local maxima to 2.9% across all data sets.

3.3.2. Goodness of cluster recovery

To examine the goodness of recovery of the groups' cluster memberships, we compare the modal clustering (i.e., assigning each group to the cluster for which the posterior classification probability is the highest) to the true clustering by means of the *Adjusted Rand Index* (*ARI*; Hubert & Arabie, 1985). The *ARI* equals 1 if the two partitions are identical, and equals 0 when the overlap between the two partitions is at chance level. The mean *ARI* over all data sets amounted to .95 (*SD* = .14) and the *ARI* was equal to 1 for no less than 86% of the data sets, which indicates a very good recovery. Out of the 1,766 solutions with at least one incorrect cluster assignment, 1,296 were in fact a local maximum. After replacing the 1,296 local maxima (obtained with 10 starts) by the solutions obtained with 50 starts (see Section 3.3.1), the number of data sets with incorrect assignments reduced from 1,766 to 912 and the overall mean of the *ARI* amounted to .98 (*SD* = .10). Table 1 presents the mean *ARI* values in function of the simulated conditions. In addition to the ones for MMG-CFA with 10 random starts, it also includes the *ARI* results after replacing the local maxima by the solutions obtained with 50 starts. When inspecting the latter results, we find

that the *ARI* was affected most by the group sizes, but also by the number and size of intercept differences and the cluster sizes (i.e., number of clusters and whether cluster sizes are equal or unequal). The interplay of these effects is depicted in Figure 1. In general, the clusters were recovered better in case of larger groups, larger or more intercept differences, less clusters or equal cluster sizes. Specifically, the *ARI* was always above .90 – indicating excellent recovery according to the guidelines by Steinley (2004) – in case of eight intercept differences of either size or two differences of .60 per cluster pair. In case of two differences of .30, the *ARI* drops well below .90 when four equal clusters are estimated for groups of 50 observations each or when four unequal clusters are estimated for group sizes of 50 or 100.

[Insert Table 1 and Figure 1 about here]

To examine the occurrence of classification uncertainty, we computed the minimum posterior probability with which a group was assigned to a cluster (according to the modal assignments), i.e., the minimum 'classification certainty' or 'CC_{min}', for each data set. For the data sets with a perfect cluster recovery (i.e., *ARI* = 1) after 10 starts, CC_{min} varied between .55 and 1.00, with a mean of .9979 (*SD* = .02). For the data sets with at least one misclassification, CC_{min} varied between .41 and 1.00 with a mean of .80 (*SD* = .19). Thus, for the simulated conditions, classification uncertainty seems to be somewhat related to misclassification.

Finally, we compared the cluster recovery of MMG-CFA to that of ML-FMM. Over all data sets, the mean *ARI* of ML-FMM amounted to .75 (*SD* = .36) and the *ARI* was equal to 1 for 65% of the data sets. For the sake of a fair comparison, we evaluated whether ML-FMM performed better with 50 rather than 10 starts for the 4,476 data sets with an incorrect clustering[2]. After replacing the results for the latter data sets with the ones after 50 starts, the mean *ARI* was .78 (*SD*

[2] Note that, for ML-FMM, we cannot trace local maxima since the true clustering for MMG-CFA cannot be used to obtain a proxy for the global maximum of ML-FMM due to the differences between the methods.

= .36). Thus, the performance of ML-FMM is clearly inferior to that of MMG-CFA when it comes to recovering the clustering that is underlying the intercept differences. Table 1 includes the mean *ARI* values for ML-FMM in function of the simulated conditions, both for 10 starts and 10 or 50 starts. It is obvious that ML-FMM performs reasonably well when it comes to picking up larger and/or more intercept differences, but that its performance drops in case of smaller or fewer intercept differences, probably because the clustering then focuses on other parameter differences.

3.3.3. Goodness of intercept recovery

To quantify how well the cluster-specific intercepts $\boldsymbol{\tau}_k$ are recovered, we calculated the mean absolute difference (*MAD*) and the root mean square error (*RMSE*) between the true and estimated intercepts:

$$MAD_{intercepts} = \frac{\sum_{k=1}^{K}\sum_{j=1}^{J}\left|\tau_{kj} - \hat{\tau}_{kj}\right|}{KJ} \quad \text{and} \quad RMSE_{intercepts} = \sqrt{\frac{\sum_{k=1}^{K}\sum_{j=1}^{J}\left(\tau_{kj} - \hat{\tau}_{kj}\right)^2}{KJ}}. \qquad (6)$$

Since different combinations of the true and estimated clusters are possible (i.e., permutational freedom of the cluster labels or 'label switching'; Tueller, Drotar, & Lubke, 2011), the estimated clusters were matched to the true clusters such that the number of misclassified groups was minimized. On average, $MAD_{intercepts}$ was equal to .03 (*SD* = .03) and $RMSE_{intercepts}$ was equal to .04 (*SD* = .04). The mean values in function of the simulated conditions, both for 10 starts and after replacing the local maxima with the best solution after 50 starts, are given in Table 2. The intercept recovery depends most on the within-cluster sample size; i.e., it improves with more groups, larger groups, less clusters and equal cluster sizes. For $RMSE_{intercepts}$, the combination of these effects is visualized in Figure 2. Clearly, recovering the intercepts is most difficult when smaller groups are assigned to four unequal clusters, especially for the case of 12 groups where the three smallest clusters are singleton clusters.

[Insert Table 2 and Figure 2 about here]

3.3.4. Goodness of loading recovery

To evaluate the recovery of the invariant factor loadings, we obtained a *goodness-of-loading-recovery statistic* (*GOLR*) by computing congruence coefficients[3] φ (Tucker, 1951) between the loadings of the true and estimated factors and averaging across factors as follows:

$$GOLR = \frac{1}{Q}\sum_{q=1}^{Q}\varphi\left(\boldsymbol{\lambda}_q, \hat{\boldsymbol{\lambda}}_q\right) \tag{7}$$

where $\boldsymbol{\lambda}_q$ and $\hat{\boldsymbol{\lambda}}_q$ indicate the true and estimated loadings of the q-th factor, respectively. The *GOLR* statistic takes values between 0 (no recovery at all) and 1 (perfect recovery). For the MMG-CFA analyses with 10 starts, the average *GOLR* is .9999 (*SD* = .00), which corresponds to an excellent recovery that is hardly affected by the manipulated conditions – probably because the total sample size is always large (i.e., at least 600).

Since *GOLR* captures the *proportional* similarity of true and estimated loadings, we also computed the *MAD* and *RMSE* between the true and estimated loadings as follows:

$$MAD_{loadings} = \frac{\sum_{q=1}^{Q}\sum_{j=1}^{J}\left|\lambda_{qj} - \hat{\lambda}_{qj}\right|}{JQ} \quad \text{and} \quad RMSE_{loadings} = \sqrt{\frac{\sum_{q=1}^{Q}\sum_{j=1}^{J}\left(\lambda_{qj} - \hat{\lambda}_{qj}\right)^2}{JQ}}. \tag{8}$$

On average, $MAD_{loadings}$ was equal to .02 (*SD* = .01) and $RMSE_{loadings}$ was equal to .03 (*SD* = .02). These values depend most on the number of groups; specifically, the mean $MAD_{loadings}$ values for 12 and 60 groups are .03 and .01, respectively, whereas for $RMSE_{loadings}$ they are .04 and .02.

3.3.5. Goodness of factor mean recovery

[3] The congruence coefficient between two column vectors **x** and **y** is defined as their normalized inner product: $\varphi(\mathbf{x},\mathbf{y}) = \mathbf{x}'\mathbf{y}\big/\sqrt{(\mathbf{x}'\mathbf{x})(\mathbf{y}'\mathbf{y})}$.

To quantify the recovery of the group- and cluster-specific factor means $\boldsymbol{\alpha}_{gk}$, we computed the *MAD* and *RMSE* between the true and estimated factor means as follows:

$$MAD_{factormeans} = \frac{\sum_{g=1}^{G} \hat{z}_{gk}^{*} \sum_{k=1}^{K} \sum_{q=1}^{Q} \left| \alpha_{gkq} - \hat{\alpha}_{gkq} \right|}{GQ} \quad \text{and}$$

$$RMSE_{factormeans} = \sqrt{\frac{\sum_{g=1}^{G} \hat{z}_{gk}^{*} \sum_{k=1}^{K} \sum_{q=1}^{Q} \left(\alpha_{gkq} - \hat{\alpha}_{gkq} \right)^{2}}{GQ}}, \tag{9}$$

where $\hat{z}_{gk}^{*}$ indicates the modal cluster memberships (i.e., assigning groups to the cluster for which $\hat{z}_{gk}$ is highest). With respect to the permutation of the cluster labels, we applied the best permutation of the estimated clusters obtained in Section 3.3.3. (i.e., minimizing misclassifications). Note that, due to the identification restrictions on the factor means per cluster, i.e., $\frac{1}{N_k} \sum_{g=1}^{G} N_g \hat{z}_{gk} \hat{\boldsymbol{\alpha}}_{gk} = \mathbf{0}$, the so-called 'alignment' (Asparouhov & Muthén, 2014) of the factor means within a cluster depends on the posterior classification probabilities for that particular cluster. Thus, misclassification and classification uncertainty cause the estimated factor means for the affected clusters to be 'misaligned' and not directly comparable to the true factor means. Therefore, we exclude the 912 data sets with an *ARI* < 1 (after 10 or 50 starts, see Section 3.3.2) and recall that classification uncertainty was rare among the data sets with an *ARI* = 1. When averaged across the 11,888 data sets with a correct clustering, $MAD_{factormeans}$ amounts to .06 (*SD* = .03) and $RMSE_{factormeans}$ amounts to .08 (*SD* = .04), which indicates a good recovery. Not surprisingly, the factor mean recovery depends most on the size of the groups; specifically, group sizes of 50, 100, 300 and 500 result in mean $MAD_{factormeans}$ values of .11, .08, .04, and .03, respectively. For $RMSE_{factormeans}$, the corresponding values are .14, .10, .06, and .04, respectively.

Additionally, we computed the correlation between the estimated and true factor means per factor and for each subset of groups that were part of the same cluster – implying that their factor means are comparable to one another – both in the estimated and true model. Note that the correlation quantifies the relative agreement between estimated and true factor means (i.e., whether the between-group differences in factor means and the rank order of the group-specific means are correctly recovered) and is insensitive to the alignment of the factor means; thus, data sets with an incorrect clustering were included. In fact, in case of a correct clustering, the group-subsets correspond to the clusters, whereas for an incorrect clustering more group-subsets than clusters are found (e.g., because a true cluster is split up in the estimated model). The mean correlation was computed across factors and group-subsets[4]. When averaged across simulated data sets, this $R_{factormeans}$ amounts to .94 (*SD* = .06). For the group sizes of 50, 100, 300 and 500, the average values are .87, .93, .97, and .98, respectively. When replacing the local maxima with the 50-start solutions, these values remain the same. For the 912 data sets with an *ARI* < 1 (and with a mean *ARI* of .68), the average $R_{factormeans}$ amounts to .90 (*SD* = .07), which still indicates a good recovery of the factor mean differences.

3.3.6. Sensitivity to local maxima and goodness-of-recovery for MMG-EFA

To evaluate whether the performance is affected by using EFA instead of CFA to model the factors, we present the overall results for MMG-EFA (with 10 starts). Out of the 12,800 solutions, 1,321 (10.3%) were found to be local maxima. The mean *ARI* was equal to .95 (*SD* = .15) and the *ARI* was equal to one for 10,996 (86%) of the data sets. The mean $MAD_{intercepts}$ amounted to .03 (*SD* = .03) and the mean $RMSE_{intercepts}$ was .05 (*SD* = .04). The mean *GOLR* was

[4] Subsets containing only one or two groups were excluded, because no correlation can be computed between one estimated and one true factor mean and because the correlation between the estimated and true factor means for two groups is always equal to 1 or −1.

.9996 (*SD* = .00), upon resolving the factors' rotational freedom by an oblique procrustes rotation of the estimated towards the true loadings. The mean $MAD_{loadings}$ and $RMSE_{loadings}$ were .01 (*SD* = .01) and .02 (*SD* = .01), respectively. Note that these values comprise the recovery of the zero loadings. The $MAD_{factormeans}$ was, on average across the 10,996 data sets with *ARI* = 1, equal to .06 (*SD* = .03), whereas the corresponding $RMSE_{factormeans}$ value was .08 (*SD* = .04). The mean $R_{factormeans}$ amounted to .94 (*SD* = .06). Thus, overall, MMG-EFA and -CFA perform equally well.

3.3.7. Model selection

For each of the 6,400 data sets generated in the first 25 replications per cell of the design, MMG-CFA analyses with one to six clusters were performed and the number of clusters with the optimal balance between log *L* and the number of free parameters *fp* was determined according to four model selection procedures (Section 2.2.3): BIC using the number of subjects *N* as the sample size (BIC_N), BIC using the number of groups *G* as the sample size (BIC_G), AIC and CHull. For BIC_N and BIC_G, the percentage of data sets for which the correct number of clusters was chosen amounts to 79.1% and 87.3%, respectively. AIC selects the correct number of clusters for 83.8% and CHull for 86.7% of the data sets. For BIC_N, the majority of the model selection mistakes pertain to the number of clusters being underestimated, whereas AIC often selects too many clusters. For BIC_G and CHull, under- and overselection of clusters occur about equally often.

The main effects of the simulated conditions on the performance of these four criteria are given in Table 3. Clearly, BIC_G and CHull show comparable performances in most conditions. The performance of the four criteria was affected most by unequal cluster sizes and the number of clusters being four rather than two and also by the group sizes and number of groups. Recall that the conditions with four unequal clusters are the ones with the most local maxima. To determine whether some model selection mistakes are due to local maxima, we repeated the model selection

for these data sets for analyses (with one to six clusters) performed with 50 starts instead of 10. Doing so avoids 341 (out of 1,336) mistakes for BIC_N, 390 (out of 816) for BIC_G, 305 (out of 1,037) for AIC and 445 (out of 851) for CHull, bringing their overall accuracy to 84.5%, 93.3%, 89.5% and 93.7%, respectively. For BIC_G and CHull, the interaction of the most important effects on this model selection performance are visualized in Figure 3. Clearly, in case of 12 groups, model selection is most difficult for four unequal clusters but improves with increasing group sizes. Interestingly, the model selection accuracy for four unequal clusters is worse for 60 groups than for 12 groups when the group sizes are 300 or 500, whereas the performance for 60 groups is superior to the one for 12 groups in the remaining conditions. This is due to a tendency to split up the large cluster in case of 60 groups (thus, when 45 groups are in the large cluster).

[Insert Table 3 and Figure 3 about here]

3.3.8. Goodness-of-recovery for too few or too many clusters

To investigate how the goodness-of-recovery is affected by selecting the wrong number of clusters, we scrutinized the recovery of the clustering and parameters for the MMG-CFA models with one to six clusters for the 6,400 data sets for which model selection was performed, excluding the ones with the correct number of clusters *K*. Of these 32,000 models, 2,118 included one or more empty clusters – clearly indicating overselection – and are thus not further evaluated in terms of recovery. When available, the solutions with 50 starts (instead of 10) were used.

Firstly, with regard to the clustering, when too few clusters are estimated, we investigated whether some of the true clusters are fused or whether the groups of (a) true cluster(s) are spread across two or more of the estimated clusters. Vice versa, when too many clusters are estimated, we evaluated whether one or more of the true clusters is split up or whether additional clusters are composed out of an amalgam of groups that belong to different true clusters. To this end, we

computed the *ARI* between the estimated clustering and the true clustering in which two true clusters were merged (in case of too few clusters) or one of the true clusters was split in two parts (in case of too many clusters). Note that all possible fusions or splits were considered, retaining the one that yielded the highest *ARI* value, and that fusions/splits of true clusters were performed until the resulting number of clusters matched the estimated one. If the resulting *ARI* equals 1, this implies that the analysis deals with the misspecification of K by cleanly fusing or splitting true clusters. Excluding models with only one cluster, this *ARI* is on average .98 (*SD* = .08), .98 (*SD* = .09), .99 (*SD* = .08), .99 (*SD* = .06), .99 (*SD* = .05), and .99 (*SD* = .05) when the estimated number of clusters is $K-2$, $K-1$, $K+1$, $K+2$, $K+3$, and $K+4$, respectively. The *ARI* equals 1 for 94.6% of the models, which indicates that the estimated clustering strongly relates to the true one in that it often merely splits or merges clusters. This depends mostly on the group sizes; i.e., *ARI* = 1 for 84.6%, 96.0%, 99.7%, and 99.5% of the models for group sizes of 50, 100, 300 and 500, respectively.

Secondly, regarding the recovery of the cluster-specific intercepts, it is important to note that: 1) underselection implies a severe information loss since too few sets of cluster-specific intercepts are estimated, and 2) overselection implies a severe misalignment of the factor means (e.g., for a true cluster that is split up) in comparison to the true solution, which also affects the intercepts. Therefore, for the models with too few clusters, we did not evaluate the intercept recovery, since for merged clusters only a weighted average of the true intercepts can be captured. To quantify the intercept recovery for too many clusters, we linked the estimated intercepts to the true ones by matching the estimated and true clusters so that they maximally agreed in terms of assignments of the groups. Thus, for example, when a true cluster was split in two, the true intercepts of this cluster are compared to those of two estimated clusters. $MAD_{intercepts}$ was on average equal to .06 (*SD* = .03), .08 (*SD* = .03), .10 (*SD* = .03), and .12 (*SD* = .04) when estimating

K+1, K+2, K+3, and K+4 clusters, respectively. The corresponding values for $RMSE_{intercepts}$ were .09 (SD = .04), .11 (SD = .04), .14 (SD = .04) and .15 (SD = .04), respectively. Probably, the intercept recovery is more contaminated by misalignment as the number of clusters increases. Optimizing the alignment based on the estimated parameters is beyond the scope of this paper (see Section 5), but we re-aligned the intercepts based on the true parameter values to somewhat unravel the recovery from misalignment. Since all true non-zero loadings have the same size, crossloadings are absent and the smallest of the true intercepts are zero for each cluster, we re-aligned the estimated intercepts per cluster and per subset of items loading on the same factor so that the smallest intercept became zero (thus, by subtracting the value of the smallest intercept). After doing so, $MAD_{intercepts}$ amounted to .04 (SD = .02), .05 (SD = .03), .05 (SD = .03), and .06 (SD = .03) for K+1, K+2, K+3, and K+4 estimated clusters, respectively, indicating a slight increase in error fitting with additional clusters. The corresponding $RMSE_{intercepts}$ values became .06 (SD = .03), .07 (SD = .03), .07 (SD = .03) and .08 (SD = .04).

Thirdly, for the factor mean recovery, $R_{factormeans}$ is inspected, which is insensitive to misalignment. On average, $R_{factormeans}$ amounts to .94 (SD = .06), .94 (SD = .06), .94 (SD = .06), .94 (SD = .07), .93 (SD = .08), .93 (SD = .08), and .92 (SD = .09) when estimating K−3, K−2, K−1, K+1, K+2, K+3, and K+4 clusters, respectively. Seemingly, the factor mean recovery is hardly affected by estimating too few or too many clusters, possibly because the intercept differences per cluster pair pertain to, at most, two out of 5 or four out of 10 items loading on the same factor. The effect of the misspecified number of clusters becomes more apparent when zooming in on the conditions with eight differences of .60 per cluster pair; i.e., $R_{factormeans}$ equals .93 (SD = .07), .93 (SD = .07), .94 (SD = .06), .93 (SD = .07), .92 (SD = .08), .91 (SD = .08), and .91 (SD = .10) when estimating K−3, K−2, K−1, K+1, K+2, K+3, and K+4 clusters, respectively.

Finally, the recovery of the invariant loadings is unaffected by under- or overselection of clusters; i.e., on average, *GOLR* amounts to .9999 (*SD* = .00), $MAD_{loadings}$ equals .02 (*SD* = .01) and $RMSE_{loadings}$ is .03 (*SD* = .02) irrespective of the number of estimated clusters.

3.4. Conclusion

Regarding the sensitivity to local maxima, we conclude that the multistart procedure of MMG-CFA with 10 starts is sufficient to largely avoid local maxima, but that it is certainly advisable to increase the number of starts in case of more clusters and/or very unequal cluster sizes. The recovery of the cluster memberships of the groups was excellent for most simulated conditions and is enhanced by larger group sizes, less clusters, equally sized clusters, or a higher cluster separability in terms of number and size of intercept differences. The cluster-specific intercepts were retrieved very well and even more so in case of more groups, larger groups, less clusters or equal cluster sizes. The recovery of the invariant factor loadings was excellent overall. With respect to the factor means, we found a good recovery that improves with a larger group size, both in terms of the factor mean *values* for all groups and the factor mean *differences* between groups within a cluster – and thus the latent mean comparisons that are of main interest to the researcher. Overall, MMG-EFA performed equally well, indicating that it is a viable option when one wants to scrutinize intercept non-invariance without imposing zero loadings that require a priori knowledge on the MM and may be overly restrictive.

For selecting the number of clusters, BIC_G and CHull were found to perform similarly, at least for the simulated conditions. BIC_G has the added value that it can automatically distinguish between one cluster (i.e., scalar invariance across all groups) and more clusters, but CHull makes no distributional assumptions and, thus, may perform better for empirical data. Furthermore, visual inspection of the CHull plot may still suggest one cluster, i.e., in case of no

clear elbow. Therefore, combining BIC_G with CHull is recommended. For CHull, one can use the free software developed by Wilderjans, Ceulemans, and Meers (2013) or the R-package that is available from https://cran.r-project.org/package=multichull. When selecting too few clusters, true clusters are merged in the estimated solution and some information on intercept differences is lost. When selecting too many clusters, true clusters are split in the estimated solution and the recovery of intercepts and factor means is affected by error fitting. When in doubt about the number of clusters, select the two or three best solutions and compare them in terms of the clusters and intercept differences. When additional clusters capture subtle intercept differences only, the more parsimonious solution should be preferred.

4. Application

In this section, we illustrate the application of MMG-FA for unraveling intercept non-invariance for data on human values measured by the 21-item Portrait Values Questionnaire (PVQ-21) during round 6 of the European Social Survey (ESS). The PVQ-21 measures 10 basic values which form four higher-order values: self-enhancement (i.e., including achievement and power), self-transcendence (i.e., benevolence and universalism), conservation (i.e., conformity, tradition, and security), and openness to change (i.e., self-direction, stimulation, and hedonism) (Schwartz, 2006). Cieciuch, Davidov, Algesheimer, and Schmidt (2018) evaluated measurement invariance of the PVQ-21 across 15 countries included in rounds 1 to 6 of the ESS: Belgium, Switzerland, Germany, Denmark, Spain, Finland, United Kingdom, Hungary, Ireland, the Netherlands, Norway, Poland, Portugal, Sweden, and Slovenia. Specifically, they tested the invariance of each higher-order value and showed that approximate invariance (Muthén & Asparouhov, 2013) could be established for openness to change and self-enhancement, whereas, for conservation and self-transcendence, this could be established only for a subset of the countries. Non-invariance

pertained mostly to the intercepts. To focus on these intercept non-invariances, we looked for a subset of countries with invariant loadings to start from. Based on the so-called 'difference output' (provided by Mplus) in the online appendices of Cieciuch et al. (2018), we inspected which countries had loadings that significantly deviated from the average. For round 6, only a few countries had significantly deviating loadings; i.e., Spain, Hungary, Portugal and Slovenia for conservation and Norway for self-transcendence. Hence, we continued with 11 countries for conservation and 14 for self-transcendence. After excluding respondents with missing values, we retained a sample size of 21,672 for the conservation data and 27,512 for self-transcendence.

To scrutinize the intercept non-invariances for the six conservation items (listed in Appendix C), we performed MMG-FA with one factor and 1 to 11 clusters (where the latter corresponds to MG-FA with group-specific intercepts). BIC_G suggests 11 clusters (Table 4), which may be an overselection (see Section 2.2.3). According to CHull, the best number of clusters is two, with a scree ratio of 2.69, but four clusters have a comparable scree ratio of 2.66. In Figure 3, we see that the CHull plot indeed has a first elbow for two clusters, but really levels off after four clusters which is why the latter solution is selected[5]. The clustering is given in Table 5. Cluster 1 contains Belgium and Sweden, whereas Switzerland, Germany, United Kingdom and Ireland are gathered in Cluster 2. Denmark and Norway make out Cluster 3 and Finland, the Netherlands, and Poland are assigned to Cluster 4. The cluster-specific intercepts and invariant factor loadings are given in Table 6. Note that, due to the identification constraints on the factor means per cluster (Section 2.2.1), the intercepts pertain to the cluster-specific item means. The intercept for the conformity item 'ipfrule' (i.e., about following rules) is higher in Cluster 1 and 2 (i.e., for Belgium,

[5] Note that, in case of two clusters, Belgium, Switzerland, Germany, United Kingdom and Ireland are in one cluster whereas Sweden, Denmark, the Netherlands, Norway, Poland and Sweden are gathered in the other cluster. In this solution, important intercept differences are overlooked, such as the ones for 'ipmodst', 'impsafe' and 'ipstrgv' between Clusters 3 and 4 (Table 6), which is why we preferred four clusters.

Sweden, Switzerland, Germany, United Kingdom, and Ireland), whereas the intercept for the tradition item ‘ipmodst’ (i.e., being humble and modest) is highest in Cluster 3 (i.e., for Denmark and Norway). The security items ‘impsafe’ and ‘ipstrgv’ (i.e., about secure surroundings and the government ensuring safety) have a higher intercept in Clusters 1 and 3 (i.e., for Belgium, Sweden, Denmark and Norway). The group- and cluster-specific means on the ‘conservation’ factor are in Table 7. These latent means are validly comparable within each cluster. For Cluster 1, we conclude that conservation is, on average, valued more in Sweden than in Belgium. In Cluster 2, it is valued most in Germany. In Cluster 3, Denmark and Norway hardly differ. In Cluster 4, the respondents of Poland value conservation the least.

[Insert Figure 4 and Tables 4, 5, 6, and 7 about here]

For the five self-transcendence items (Appendix C), we performed MMG-FA with one factor and 1 to 14 clusters (where the latter implies group-specific intercepts). BIC_G indicates 14 clusters (Table 4), again likely to be an overselection, whereas CHull selects three clusters, with a scree ratio of 2.32 and a clear elbow in Figure 3. The cluster assignments are given in Table 5. Specifically, Belgium, Switzerland, Germany, Denmark are assigned to Cluster 1. Spain, United Kingdom, Ireland, the Netherlands, Portugal, and Sweden are in Cluster 2. Hungary, Poland, and Slovenia make out Cluster 3. Table 8 contains the factor loadings and cluster-specific intercepts. The intercept for the benevolence item ‘iplylfr’ (i.e., about loyalty to friends) is highest in Cluster 2 and lowest in Cluster 1. The intercept for the universalism item ‘impenv’ (i.e., about looking after the environment) is highest in Cluster 2 and lowest in Cluster 3. For the other items, the differences are very subtle. The group- and cluster-specific factor means (Table 9) indicate that, in Cluster 1, the respondents from Switzerland value self-transcendence the least, whereas, in

Clusters 2 and 3, this is the case for Spain and Slovenia, respectively. Also, in Cluster 2, self-transcendence is valued notably more by the respondents from Portugal.

[Insert Tables 8 and 9 about here]

We conclude that the intercept non-invariances for conservation were scrutinized by assigning the 11 countries to four clusters and inspecting four sets of cluster-specific intercepts, whereas, for self-transcendence, the non-invariances across 14 countries were captured by three clusters and three sets of cluster-specific intercepts. This is clearly more efficient and insightful than making 55 and 91 pairwise comparisons of country-specific intercepts, respectively. Based on the MMG-FA solutions, latent mean comparisons could be made per cluster.

5. Discussion

Mixture multigroup factor analysis is an efficient and insightful way to trace measurement non-invariances across many groups. This paper focused on comparing intercepts across groups for which metric invariance holds but scalar invariance fails. Specifically, groups with (near-)identical intercepts end up in the same mixture cluster and are modeled with one set of cluster-specific intercepts. Since other measurement or structural parameters are either invariant or allowed to differ across groups, the clustering is only affected by the intercepts. As a result, MMG-FA not only drastically reduces the number of intercepts to compare but also takes an essential step towards finding clusters of groups for which latent mean comparisons are valid. Indeed, the resulting factor means can be compared among groups in the same cluster.

When one aims to make valid comparisons across all groups, instead of per cluster, the MMG-FA solution provides the user with useful clues on how to continue. Firstly, the comparison of the cluster-specific intercepts may indicate that one or a few items are causing the non-invariance and, thus, that excluding these items – or making their intercepts non-invariant to

continue with partial invariance (Byrne, Shavelson, & Muthén, 1989) – allows for valid comparisons across all groups in the data set. Secondly, the clustering may identify a few problematic groups and, then, excluding these groups is an option. Thirdly, a combination of non-invariant items and groups may be found. In any case, one should consider which (combination of) items or groups to remove based on substantive considerations and the amount of retained data.

The cluster-specific intercepts – and the differences between them – depend on the 'alignment' of intercepts and factor means within each cluster. The identification restrictions imposed on the factor means per cluster correspond to one particular 'alignment'. Specifically, when latent means for the groups in one cluster are higher, overall, than in another cluster, this would be captured by the intercepts rather than the factor means. Thus, in order to optimally compare intercepts across clusters (i.e., without finding differences that are actually due to latent mean differences), the cluster-specific solutions should be aligned to maximize the between-cluster agreement of intercepts. Currently, in case of no crossloadings and no (or hardly any) classification uncertainty, one can take the MMG-FA clusters and use them as the groups in multigroup factor alignment (Asparouhov & Muthén, 2014), imposing invariant loadings. In the future, MMG-FA will be extended with an alignment that accommodates crossloadings.

To explore predictors of the non-invariances captured by the clustering, MMG-FA can be extended to allow for group-level covariates explaining the cluster memberships (Lubke & Muthén, 2005, 2007). Alternatively, covariates can be added after estimating MMG-FA by using the three-step approach (Vermunt, 2010). This would enable researchers, for instance, to use economic, political or cultural indicators to explain between-country differences in intercepts.

Since partial invariance of loadings does not preclude latent mean comparisons (Byrne, Shavelson, & Muthén, 1989), MMG-FA should be able to build on partial metric invariance.

Therefore, in the future, it will be extended to allow for loadings that are partially group-specific. For the time being, one could remove items with non-invariant loadings before applying MMG-FA for tracing intercept non-invariances and potentially combine it with the standard multigroup factor analysis with partially invariant loadings and group-specific intercepts.

When one aims for latent mean comparisons but the loadings are not (fully or partially) invariant, one can first use the previously presented variant of MMG-FA (De Roover, Vermunt, & Ceulemans, in press) to obtain clusters of groups with invariant loadings. As a second step, one can apply MMG-FA for intercept invariance within each of those clusters. A MMG-FA variant that simultaneously deals with loading and intercept non-invariances will be considered in the future. Note that such a method needs to be thoroughly evaluated since it could become less effective when the underlying group-clusters for the loadings are very different from the ones for the intercepts. This would result either in a very high number of clusters when capturing both types of differences at the same time or – when the model selection fails to select this high number of clusters – a few clusters that fail to uncover any of the measurement model differences properly.

Finally, exact and full invariance of the intercepts within a cluster may be too restrictive and unrealistic, especially in case of many groups. While awaiting MMG-FA extensions capturing partial or approximate invariance within the clusters, one can apply existing approaches – such as modification indices (Sörbom, 1989), item-deletion strategies (Byrne & van de Vijver, 2010) and Bayesian structural equation modeling (Muthén & Asparouhov, 2013) – per MMG-FA cluster.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Pet rov & F. Caski (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

Anderson, J. C., & Gerbing, D. W. (1982). Some methods for respecifying measurement models to obtain unidimensional construct measurement. *Journal of Marketing Research, 19(4),* 453-460.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: a multidisciplinary journal, 16(3),* 397-438.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21,* 495-508.

Beierlein, C., Kuntz, A., & Davidov, E. (2016). Universalism, conservation and attitudes toward minority groups. *Social Science Research, 58,* 68-79.

Boer, D., Hanke, K., & He, J. (2018). On detecting systematic measurement error in cross-cultural research: A review and critical reflection on equivalence and invariance tests. *Journal of Cross-Cultural Psychology, 49(5),* 713-734.

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research, 36,* 48-86.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111-150.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement equivalence. *Psychological Bulletin, 105,* 456–466.

Byrne, B. M., & Van de Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10(2),* 107-132.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1(2),* 245-276.

Ceulemans, E., & Kiers, H. A. (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.

Ceulemans, E., & Van Mechelen, I. (2005). Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection. *Psychometrika, 70(3),* 461-480.

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.

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9(2),* 233-255.

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for approximate measurement invariance of human values in the European Social Survey. *Sociological Methods & Research, 47(4),* 665-686.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology, 43(4),* 558-575.

De Roover, K., & Vermunt, J. K. (2019). On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. *Structural Equation Modeling: A Multidisciplinary Journal, 26,* 905-923.

De Roover, K., Vermunt, J., & Ceulemans, E. (in press). Mixture multigroup factor analysis for unraveling factor loading non-invariance across many groups. *Psychological Methods.*

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling, 16,* 295-314.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272-299.

Gerbing, D. W., & Hamilton, J. G. (1996). Viability of exploratory factor analysis as a precursor to confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 3,* 62-72.

Greiff, S., & Scherer, R. (2018). Still Comparing Apples With Oranges? Some thoughts on the principles and practices of measurement invariance testing. *European Journal of Psychological Assessment, 34,* 141-144.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7,* 534-556.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Inglehart, R. F., Basanez, M., Basanez, M., & Moreno, A. (1998). Human values and beliefs: A cross-cultural sourcebook. *University of Michigan Press.*

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

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

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.

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician, 12,* 209–229.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10(1),* 21-39.

Lubke, G., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling, 14(1),* 26-47.

Lukočienė, O., Varriale, R., & Vermunt, J. K. (2010). The simultaneous decision(s) about the number of lower-and higher-level classes in multilevel latent class analysis. *Sociological Methodology, 40(1),* 247-283.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111(3),* 490-504.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.

McCrae, R. R., Zonderman, A. B., Costa Jr, P. T., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the Revised NEO Personality Inventory: Confirmatory

factor analysis versus Procrustes rotation. *Journal of Personality and Social Psychology, 70(3),* 552-566.

McLachlan, G., & Peel, D. (2000). *Finite Mixture Models.* New York, NY: Wiley.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17,* 313-335.

Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus web notes, 17,* 1-48.

Muthén, L. K., & Muthén, B. O. (2005). *Mplus: Statistical analysis with latent variables: User's guide.* Los Angeles, CA: Muthén & Muthén.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14(4),* 535-569.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461-464.

Schwartz, S. (2006). A theory of cultural value orientations: Explication and applications. *Comparative Sociology, 5,* 137-182.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Sörbom, D. (1989). Model modification. *Psychometrika, 54(3),* 371–384.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91,* 1292–1306.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods, 9,* 386-396.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling: a Multiisciplinary Journal, 20(4),* 640-657.

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling, 18(1),* 110-131.

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research, 47(2),* 247-275.

Vermunt, J. K. (2010). Latent Class Modeling with Covariates: Two Improved Three-Step Approaches. *Political Analysis, 18,* 450–469.

Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2016). *Upgrade Manual for Latent GOLD 5.1.* Belmont, MA: Statistical Innovations Inc.

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2013). CHull: A generic convex hull based model selection method. *Behavior Research Methods, 45,* 1-15.

Table 1. Mean Adjusted Rand Index (*ARI*) of the estimated clustering of the groups in comparison to the true clustering, in function of the simulated conditions, for MMG-CFA (left) with 10 random starts and with 10 or 50 starts (i.e., 50 starts when 10 starts resulted in a local maximum), and for ML-FMM (right) with 10 starts and with 10 or 50 starts (i.e., 50 starts when *ARI* was < 1 after 10 starts).

	ARI MMG-CFA		*ARI* ML-FMM	
	10 starts	10 or 50 starts	10 starts	10 or 50 starts
$G = 12$	.97	.98	.73	.74
$G = 60$	.94	.98	.78	.81
$N_g = 50$	.91	.94	.70	.72
$N_g = 100$	.95	.98	.75	.77
$N_g = 300$	.97	.99	.78	.80
$N_g = 500$	.98	.99	.79	.81
$K = 2$	.99	.99	.87	.87
$K = 4$	.91	.96	.63	.68
equal clusters	.99	.99	.86	.86
unequal clusters	.91	.96	.65	.69
$Q = 2$	.95	.98	.79	.81
$Q = 4$	.95	.98	.72	74
intercept differences of .60	.97	.99	.90	.94
intercept differences of .30	.93	.96	.60	.62
8 intercept differences	.97	.99	.90	.94
2 intercept differences	.93	.96	.60	.61
overall	.95	.98	.75	.78

Table 2. Mean Absolute Difference (*MAD*) and Root Mean Square Error (*RMSE*) between the estimated and true cluster-specific intercepts, in function of the simulated conditions, for MMG-CFA with 10 random starts and 10 or 50 random starts (i.e., 50 starts when 10 starts resulted in a local maximum).

	$MAD_{intercepts}$		$RMSE_{intercepts}$	
	MMG-CFA 10 starts	MMG-CFA 10 or 50 starts	MMG-CFA 10 starts	MMG-CFA 10 or 50 starts
$G = 12$	.04	.04	.05	.05
$G = 60$	.03	.02	.04	.03
$N_g = 50$	.05	.05	.07	.06
$N_g = 100$	.04	.03	.05	.04
$N_g = 300$	.02	.02	.03	.03
$N_g = 500$	.02	.02	.03	.02
$K = 2$	.02	.02	.03	.03
$K = 4$	.04	.04	.06	.05
equal clusters	.02	.02	.03	.03
unequal clusters	.04	.03	.06	.05
$Q = 2$	.03	.03	.04	.04
$Q = 4$	.03	.03	.05	.04
intercept differences of .60	.03	.03	.05	.04
intercept differences of .30	.03	.03	.04	.04
8 intercept differences	.03	.03	.05	.04
2 intercept differences	.03	.03	.04	.04
overall	.03	.03	.04	.04

Table 3. Percentage of data sets for which BIC_N, BIC_G, AIC and CHull select the correct number of clusters *K*, in function of the simulated conditions.

	BIC_N	BIC_G	AIC	CHull
$G = 12$	77.8	92.0	92.2	91.3
$G = 60$	80.4	82.5	75.4	82.1
$N_g = 50$	64.5	79.4	76.2	80.5
$N_g = 100$	75.2	87.3	83.3	87.1
$N_g = 300$	88.2	90.9	87.7	89.4
$N_g = 500$	88.6	91.4	88.1	89.8
$K = 2$	95.3	99.3	93.8	98.7
$K = 4$	62.9	75.2	73.8	74.7
equal clusters	91.6	98.2	93.3	97.5
unequal clusters	66.6	76.3	74.3	75.9
$Q = 2$	78.9	87.3	84.7	86.9
$Q = 4$	79.3	87.2	82.9	86.5
intercept differences of .60	87.4	90.6	87.1	89.3
intercept differences of .30	70.8	83.9	80.5	84.1
8 intercept differences	87.4	90.1	86.8	89.1
2 intercept differences	70.8	84.4	80.8	84.3
overall	79.1	87.3	83.8	86.7

Table 4. Loglikelihood (log *L*), number of free parameters (fp), BIC_G and CHull scree ratio for MMG-CFA models for the ESS human values data, with 1 to 11 clusters for conservation (above) and with 1 to 14 clusters for self-transcendence (below), where the latter models imply group-specific intercepts. For each criterion, the value for the best model (or the two best models) is in bold face.

Number of clusters	log *L*	fp	BIC_G	CHull scree ratio
K = 1	-202976.2	98	406187.3	0.00
K = 2	-201886.0	104	404021.4	**2.69**
K = 3	-201480.8	110	403225.4	1.71
K = 4	-201244.0	116	402766.2	**2.66**
K = 5	-201155.0	122	402602.5	1.01
K = 6	-201066.9	128	402440.8	1.48
K = 7	-201007.5	134	402336.3	1.06
K = 8	-200951.3	140	402238.3	1.15
K = 9	-200902.5	146	402155.1	2.50
K = 10	-200883.0	152	402130.4	1.17
K = 11	-200866.3	158	**402111.4**	0.00
Number of clusters	log *L*	fp	BIC_G	CHull scree ratio
K = 1	-172820.7	106	345921.1	/
K = 2	-172461.9	111	345216.8	1.36
K = 3	-172198.1	116	344702.4	**2.32**
K = 4	-172084.5	121	344488.2	1.30
K = 5	-171996.9	126	344326.3	1.10
K = 6	-171917.5	131	344180.7	1.93
K = 7	-171876.3	136	344111.5	1.30
K = 8	-171844.6	141	344061.4	1.12
K = 9	-171816.3	146	344017.9	1.10
K = 10	-171790.6	151	343979.8	1.54
K = 11	-171774.0	156	343959.7	1.34
K = 12	-171761.6	161	343948.0	1.18
K = 13	-171751.1	166	343940.2	1.54
K = 14	-171744.3	171	**343939.8**	/

Table 5. Clustering of MMG-FA models for the ESS human values data on conservation and self-transcendence. All countries are assigned to the clusters with posterior probability $\hat{z}_{gk} = 1$.

Country	Clustering for conservation	Clustering for self-transcendence
Belgium	1	1
Switzerland	2	1
Germany	2	1
Denmark	3	1
Spain	/	2
Finland	4	1
United Kingdom	2	2
Hungary	/	3
Ireland	2	2
Netherlands	4	2
Norway	3	/
Poland	4	3
Portugal	/	2
Sweden	1	2
Slovenia	/	3

Table 6. Invariant loadings and cluster-specific intercepts of the MMG-FA model for the six conservation items. Intercepts that are at least .30 higher than in one of the other clusters are indicated in bold face.

Item	Loadings Conservation	Intercepts Cluster 1	Intercepts Cluster 2	Intercepts Cluster 3	Intercepts Cluster 4
ipfrule	0.76	**3.21**	**3.28**	2.82	2.78
ipbhprp	0.80	2.75	2.67	2.55	2.66
ipmodst	0.51	2.55	2.56	**3.30**	**2.86**
imptrad	0.66	2.83	2.73	2.85	2.62
impsafe	0.66	**2.64**	2.21	**2.86**	2.35
ipstrgv	0.59	**2.61**	2.19	**2.74**	2.35

Table 7. Group- and cluster-specific means on the ‘conservation’ factor, for each country and for the cluster that country is assigned to.

Country	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Belgium	-0.28			
Switzerland		0.03		
Germany		0.17		
Denmark			0.03	
Finland				0.25
United Kingdom		-0.12		
Ireland		-0.11		
Netherlands				0.30
Norway			-0.03	
Poland				-0.58
Sweden	0.29			

Table 8. Invariant loadings and cluster-specific intercepts of the MMG-FA model for the five self-transcendence items. Intercepts that are at least .30 higher than in one of the other clusters are indicated in bold face.

Item	Loadings Self-Transcendence	Intercepts Cluster 1	Intercepts Cluster 2	Intercepts Cluster 3
iphlppl	0.58	2.00	2.05	2.09
iplylfr	0.50	1.66	1.93	1.86
ipeqopt	0.45	1.98	1.93	1.84
ipudrst	0.54	2.17	2.34	2.30
impenv	0.49	2.01	**2.16**	1.82

Table 9. Group- and cluster-specific means on the ‘self-transcendence’ factor, for each country and for the cluster that country is assigned to.

Country	Cluster 1	Cluster 2	Cluster 3
Belgium	0.11		
Switzerland	-0.17		
Germany	-0.06		
Denmark	-0.01		
Spain		-0.57	
Finland	0.11		
United Kingdom		-0.14	
Hungary			0.13
Ireland		-0.03	
Netherlands		0.13	
Poland			0.08
Portugal		0.71	
Sweden		-0.16	
Slovenia			-0.32

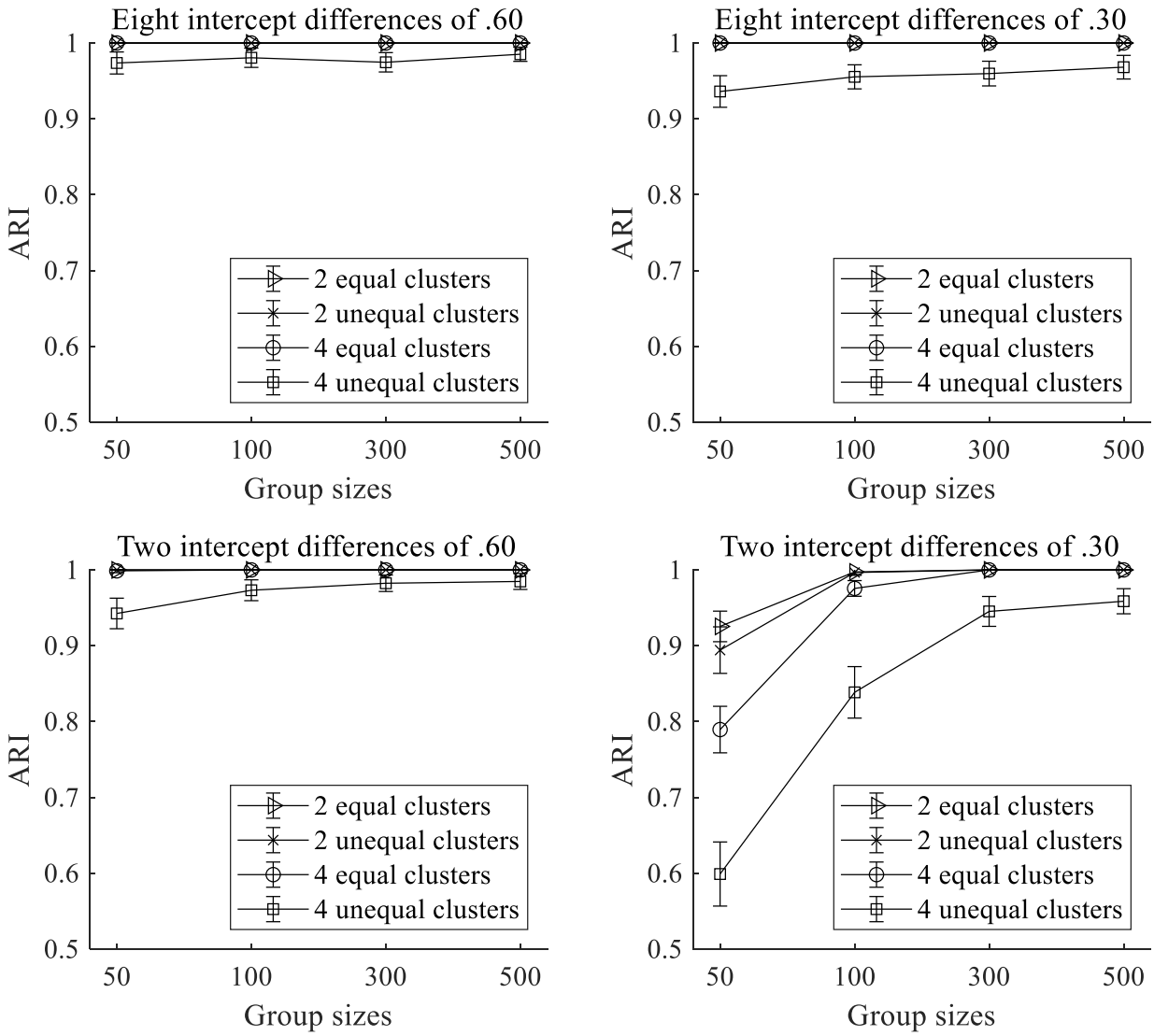

Figure 1. Mean adjusted Rand index (*ARI*) and associated 95% confidence intervals in function of the group sizes, number of clusters, cluster sizes, and number and size of intercept differences, for MMG-CFA with 10 or 50 starts (i.e., 50 starts when 10 starts resulted in a local maximum).

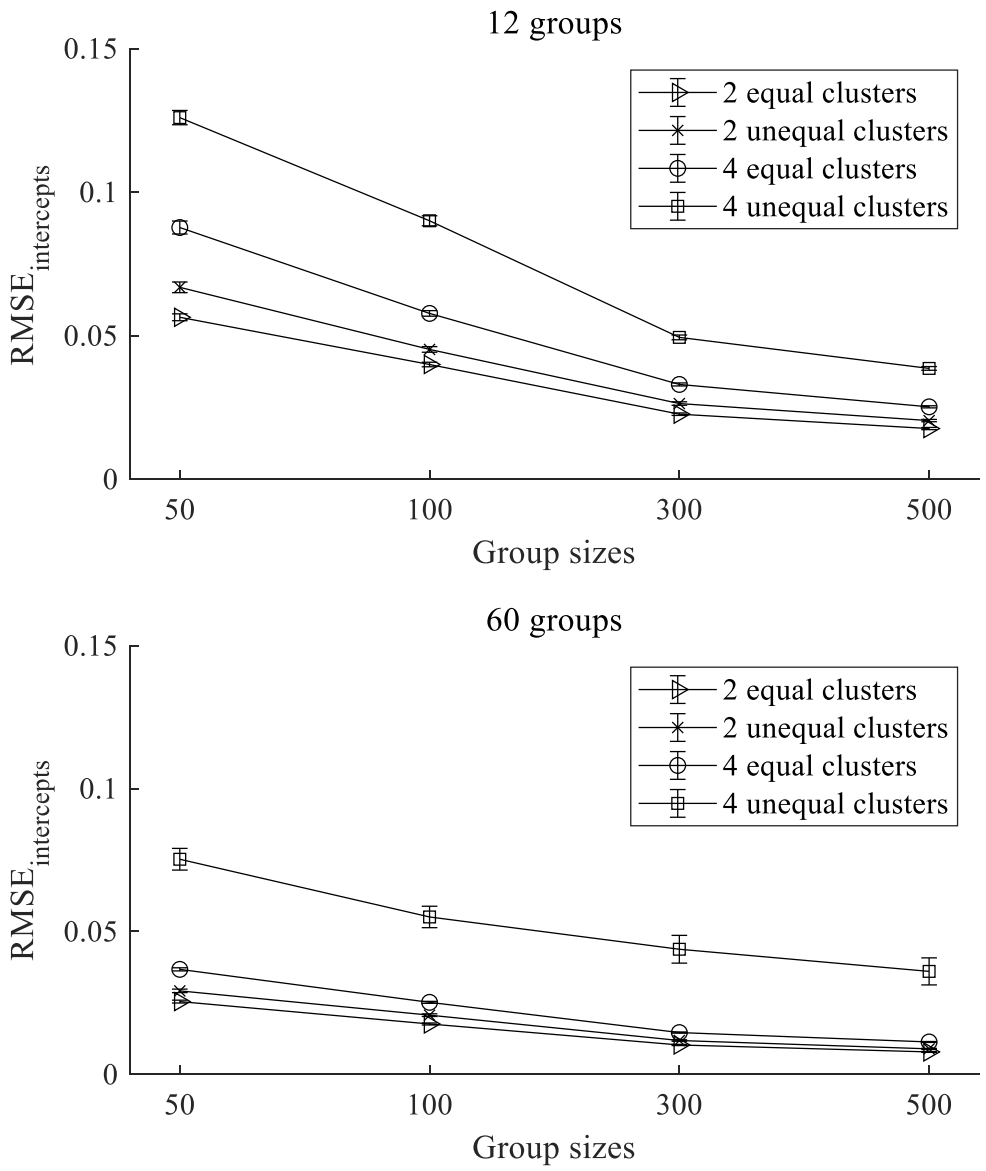

Figure 2. Root Mean Square Error (*RMSE*) and associated 95% confidence intervals in function of the number of groups, group sizes, number of clusters, and cluster sizes, for MMG-CFA with 10 or 50 starts (i.e., 50 starts when 10 starts resulted in a local maximum).

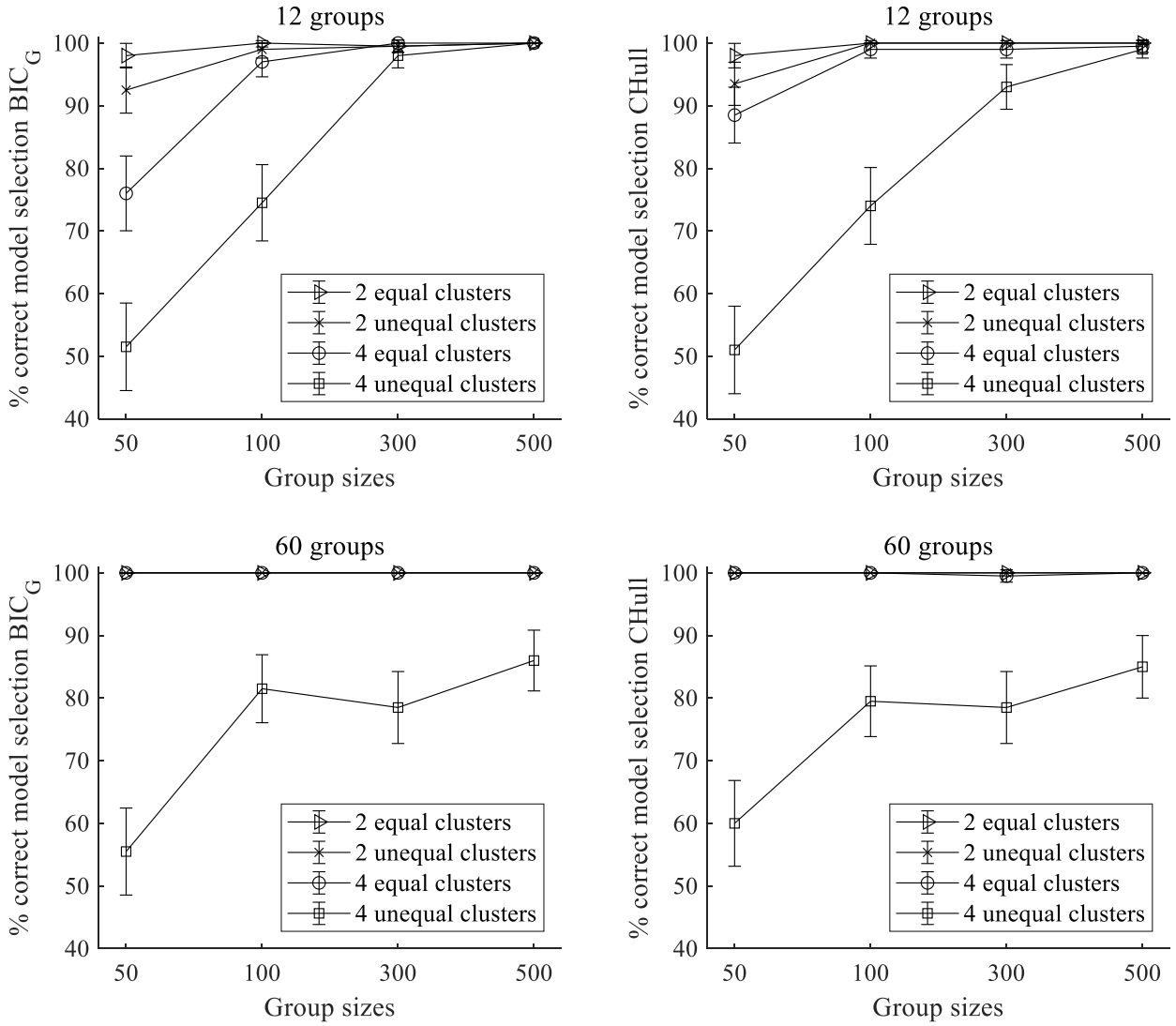

Figure 3. Percentage of data sets for which BIC_G (left) and CHull (right) select the correct number of clusters *K* and associated 95% confidence intervals in function of the number of groups, group sizes, number of clusters, and cluster sizes, for MMG-CFA with 10 or 50 starts (i.e., 50 starts when 10 starts resulted in a model selection error).

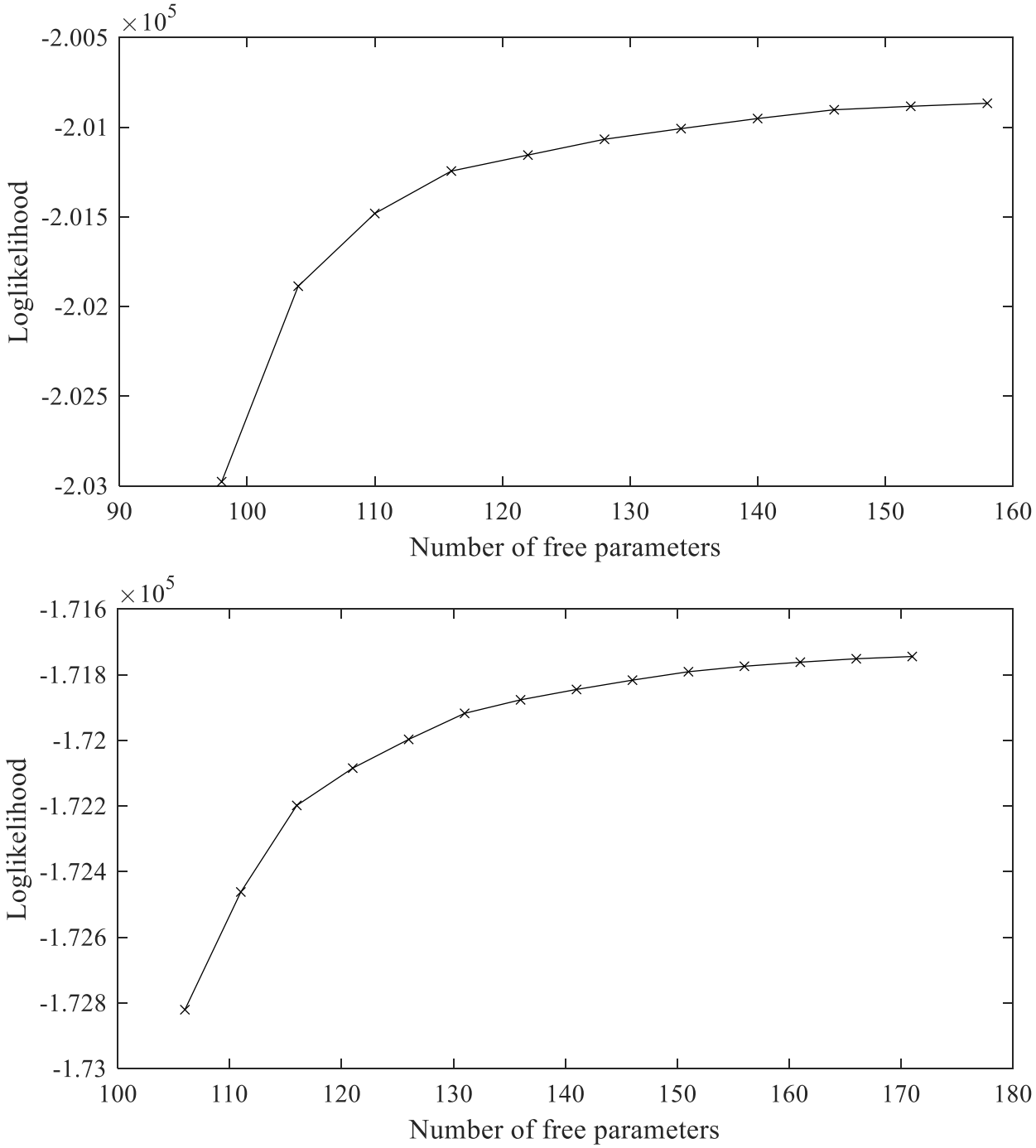

Figure 4. Convex Hull (CHull) plot of the loglikelihood in function of the number of free parameters for MMG-FA models for the ESS human values data, with 1 to 11 clusters for conservation (above) and with 1 to 14 clusters for self-transcendence (below), where the latter models correspond to MG-FA with group-specific intercepts.

Appendix A: AECM algorithm and multistart procedure

As in all mixture models, log L (Equation 4) – also referred to as the 'observed-data loglikelihood' – is complicated by the latent clustering of the groups, making it hard to maximize log L directly. Therefore, the EM algorithm makes use of the so-called 'complete-data loglikelihood', i.e., the loglikelihood when the latent (or 'missing') data is assumed to be known (i.e., the joint distribution of the observed and latent data). For estimating MMG-FA as specified in Section 2.2.1, we make use of an alternating expectation conditional maximization (AECM) algorithm (Meng, & Van Dyk, 1997), where two cycles are alternated. In cycle 1, only the groups' cluster memberships z_{gk} are considered as missing data, which leads to the following $\log L_c^1$:

$$
\begin{aligned}
\log L_c^1 &= \log\left[\prod_{k=1}^{K}\prod_{g=1}^{G}\prod_{n_g=1}^{N_g}\left(f\left(z_{gk}\right)f\left(\mathbf{x}_{n_g}\mid z_{gk}\right)\right)\right] \\
&= \log\left[\prod_{k=1}^{K}\prod_{g=1}^{G}\left(\pi_k\prod_{n_g=1}^{N_g}MVN\left(\mathbf{x}_{n_g};\boldsymbol{\tau}_k+\boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk},\boldsymbol{\Sigma}_g\right)\right)^{z_{gk}}\right] \\
&= \sum_{g=1}^{G}\sum_{k=1}^{K}z_{gk}\left(\log\left(\pi_k\right)+\sum_{n_g=1}^{N_g}\log\left(\frac{1}{\left(2\pi\right)^{J/2}\left|\boldsymbol{\Sigma}_g\right|^{1/2}}\exp\left(-\frac{1}{2}\tilde{\mathbf{x}}'_{n_gk}\boldsymbol{\Sigma}_g^{-1}\tilde{\mathbf{x}}_{n_gk}\right)\right)\right) \\
&= \sum_{g=1}^{G}\sum_{k=1}^{K}z_{gk}\log\left(\pi_k\right)-\frac{NJ}{2}\log\left(2\pi\right)-\frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Sigma}_g\right|\right)-\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}z_{gk}\left(\sum_{n_g=1}^{N_g}\left(tr\left(\tilde{\mathbf{x}}_{n_gk}\tilde{\mathbf{x}}'_{n_gk}\boldsymbol{\Sigma}_g^{-1}\right)\right)\right)
\end{aligned}
\tag{10}
$$

where $\tilde{\mathbf{x}}_{n_gk}=\mathbf{x}_{n_g}-\boldsymbol{\tau}_k-\boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk}$ and $\boldsymbol{\Sigma}_g=\boldsymbol{\Lambda}\boldsymbol{\Phi}_g\boldsymbol{\Lambda}'+\boldsymbol{\Psi}_g$. By inserting the expected values of the missing cluster memberships (Equation 11), also referred to as the posterior classification probabilities, the expected value of $\log L_c^1$ is obtained (Equation 12):

$$
\hat{z}_{gk}=E\left[z_{gk}=1\mid\mathbf{X}_g\right]=\frac{\pi_k f_{gk}\left(\mathbf{X}_g;\boldsymbol{\theta}_{gk}\right)}{\sum_{k'=1}^{K}\pi_{k'}f_{gk'}\left(\mathbf{X}_g;\boldsymbol{\theta}_{gk'}\right)}, \tag{11}
$$

$$E\left[\log L_c^1\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}\log\left(\pi_k\right) - \frac{NJ}{2}\log\left(2\pi\right) - \frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Sigma}_g\right|\right) - \frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}N_{gk}tr\left(\mathbf{S}_{gk}\boldsymbol{\Sigma}_g^{-1}\right). \quad (12)$$

Note that, in Equation 12, $N_{gk} = \hat{z}_{gk}N_g$ and $\mathbf{S}_{gk} = \frac{1}{N_g}\sum_{n_g=1}^{N_g}\tilde{\mathbf{x}}_{n_gk}\tilde{\mathbf{x}}'_{n_gk}$. By taking the derivative of $E\left[\log L_c^1\right]$ and equating it to zero – using the Lagrange multiplier method for imposing $\sum_{k=1}^{K}\hat{\pi}_k = 1$ – we obtain the following parameter updates (i.e., conditional maximization steps) for cycle 1:

$$\hat{\pi}_k = \frac{1}{G}\sum_{g=1}^{G}\hat{z}_{gk}, \quad (13)$$

$$\hat{\boldsymbol{\tau}}_k = \left(\sum_{g=1}^{G}N_{gk}\boldsymbol{\Sigma}_g^{-1}\right)^{-1}\left(\sum_{g=1}^{G}N_{gk}\boldsymbol{\Sigma}_g^{-1}\left(\bar{\mathbf{x}}_g - \boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk}\right)\right), \quad (14)$$

$$\hat{\boldsymbol{\alpha}}_{gk} = \left(\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Sigma}_g^{-1}\left(\bar{\mathbf{x}}_g - \boldsymbol{\tau}_k\right). \quad (15)$$

where $\bar{\mathbf{x}}_g$ is the observed mean for group g.

In cycle 2, both the cluster memberships z_{gk} and the factor scores $\boldsymbol{\eta}_{n_gk}$ are considered as missing data and the complete-data loglikelihood becomes:

$$\log L_c^2 = \log\left[\prod_{k=1}^{K}\prod_{g=1}^{G}\prod_{n_g=1}^{N_g}\left(f\left(z_{gk}\right)f\left(\boldsymbol{\eta}_{n_{gk}} \mid z_{gk}\right)f\left(\mathbf{x}_{n_g} \mid \boldsymbol{\eta}_{n_{gk}}, z_{gk}\right)\right)\right]$$

$$= \log\left[\prod_{k=1}^{K}\prod_{g=1}^{G}\left(\pi_k\prod_{n_g=1}^{N_g}\left(MVN\left(\boldsymbol{\eta}_{n_gk};\boldsymbol{\alpha}_{gk},\boldsymbol{\Phi}_g\right)MVN\left(\mathbf{x}_{n_g};\boldsymbol{\tau}_k+\boldsymbol{\Lambda}\boldsymbol{\eta}_{n_gk},\boldsymbol{\Psi}_g\right)\right)\right)^{z_{gk}}\right]$$

$$= \sum_{g=1}^{G}\sum_{k=1}^{K} z_{gk}\left(\begin{array}{l}\log\left(\pi_k\right)+\sum_{n_g=1}^{N_g}\log\left(\frac{1}{\left(2\pi\right)^{Q/2}\left|\boldsymbol{\Phi}_g\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{\eta}_{n_gk}-\boldsymbol{\alpha}_{gk}\right)'\boldsymbol{\Phi}_g^{-1}\left(\boldsymbol{\eta}_{n_gk}-\boldsymbol{\alpha}_{gk}\right)\right)\right)\\ +\sum_{n_g=1}^{N_g}\log\left(\frac{1}{\left(2\pi\right)^{J/2}\left|\boldsymbol{\Psi}_g\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k-\boldsymbol{\Lambda}\boldsymbol{\eta}_{n_gk}\right)'\boldsymbol{\Psi}_g^{-1}\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k-\boldsymbol{\Lambda}\boldsymbol{\eta}_{n_gk}\right)\right)\right)\end{array}\right)$$

$$= \sum_{g=1}^{G}\sum_{k=1}^{K} z_{gk}\log\left(\pi_k\right)-\frac{N\left(J+Q\right)}{2}\log\left(2\pi\right)-\frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Phi}_g\right|\right)-\frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Psi}_g\right|\right)$$

$$-\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K} z_{gk}\left(\sum_{n_g=1}^{N_g}\left(\begin{array}{l}tr\left(\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'\boldsymbol{\Phi}_g^{-1}\right)+tr\left(\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k\right)\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k\right)'\boldsymbol{\Psi}_g^{-1}\right)-\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k\right)'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\tilde{\boldsymbol{\eta}}_{n_gk}\\ -\tilde{\boldsymbol{\eta}}_{n_gk}'\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\left(\mathbf{x}_{n_g}-\boldsymbol{\tau}_k\right)+tr\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'\right)\end{array}\right)\right) \quad (16)$$

where $\tilde{\boldsymbol{\eta}}_{n_gk} = \boldsymbol{\eta}_{n_gk} - \boldsymbol{\alpha}_{gk}$. The expected values of z_{gk}, $\tilde{\boldsymbol{\eta}}_{n_gk}$ and $\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'$ are inserted in Equation 16, which yields the following expected value of $\log L_c^2$:

$$E\left[\log L_c^2\right] = \sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}\log\left(\pi_k\right)-\frac{N\left(J+Q\right)}{2}\log\left(2\pi\right)-\frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Phi}_g\right|\right)$$

$$-\frac{1}{2}\sum_{g=1}^{G}N_g\log\left(\left|\boldsymbol{\Psi}_g\right|\right)-\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}tr\left(\boldsymbol{\Phi}_g^{-1}\sum_{n_g=1}^{N_g}E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'\right]\right)-\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}N_{gk}tr\left(\mathbf{S}_{gk}\boldsymbol{\Psi}_g^{-1}\right)$$

$$+\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}tr\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\tilde{\mathbf{x}}_{n_gk}\sum_{n_g=1}^{N_g}E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\right]'\right)+\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}tr\left(\tilde{\mathbf{x}}_{n_gk}'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\sum_{n_g=1}^{N_g}E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\right]\right) \quad (17)$$

$$-\frac{1}{2}\sum_{g=1}^{G}\sum_{k=1}^{K}\hat{z}_{gk}tr\left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\sum_{n_g=1}^{N_g}E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'\right]\right)$$

where $\hat{z}_{gk}$ corresponds to the posterior classification probability (Equation 11) and $E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\right]$ and $E\left[\tilde{\boldsymbol{\eta}}_{n_gk}\tilde{\boldsymbol{\eta}}_{n_gk}'\right]$ are equal to (McLachlan & Krishnan, 2007):

$$E\left[\tilde{\boldsymbol{\eta}}_{n_g k}\right] = E\left[\tilde{\boldsymbol{\eta}}_{n_g k} \mid \mathbf{x}_{n_g}, z_{gk} = 1\right] = \boldsymbol{\beta}_g \tilde{\mathbf{x}}_{n_g k} \quad \textit{with} \quad \boldsymbol{\beta}_g = \boldsymbol{\Phi}_g \boldsymbol{\Lambda}' \left(\boldsymbol{\Lambda}\boldsymbol{\Phi}_g \boldsymbol{\Lambda}' + \boldsymbol{\Psi}_g\right)^{-1}, \tag{18}$$

$$E\left[\tilde{\boldsymbol{\eta}}_{n_g k}\tilde{\boldsymbol{\eta}}'_{n_g k}\right] = E\left[\tilde{\boldsymbol{\eta}}_{n_g k}\tilde{\boldsymbol{\eta}}'_{n_g k} \mid \mathbf{x}_{n_g}, z_{gk} = 1\right] = \boldsymbol{\Phi}_g - \boldsymbol{\beta}_g \boldsymbol{\Lambda}\boldsymbol{\Phi}_g + \boldsymbol{\beta}_g \tilde{\mathbf{x}}_{n_g k}\tilde{\mathbf{x}}'_{n_g k}\boldsymbol{\beta}'_g. \tag{19}$$

where $\tilde{\mathbf{x}}_{n_g k} = \mathbf{x}_{n_g k} - \boldsymbol{\mu}_{gk}$ and $\boldsymbol{\mu}_{gk} = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk}$. By taking the derivative of $E\left[\log L_c^2\right]$ and equating it to zero – again imposing $\sum_{k=1}^{K} \hat{\pi}_k = 1$ – we obtain the conditional maximization steps for cycle 2:

$$\hat{\pi}_k = \frac{1}{G}\sum_{g=1}^{G} \hat{z}_{gk} \tag{20}$$

$$\hat{\boldsymbol{\tau}}_k = \left(\sum_{g=1}^{G} N_{gk}\boldsymbol{\Psi}_g^{-1}\right)^{-1}\left(\sum_{g=1}^{G} N_{gk}\boldsymbol{\Psi}_g^{-1}\left(\bar{\mathbf{x}}_g - \boldsymbol{\Lambda}\boldsymbol{\alpha}_{gk} - \boldsymbol{\Lambda}\boldsymbol{\beta}_g\left(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_{gk}\right)\right)\right) \tag{21}$$

$$\hat{\boldsymbol{\alpha}}_{gk} = \left(\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}_g^{-1}\left(\bar{\mathbf{x}}_g - \boldsymbol{\tau}_k - \boldsymbol{\Lambda}\boldsymbol{\beta}_g\left(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_{gk}\right)\right) \tag{22}$$

$$\hat{\boldsymbol{\lambda}}_j^{nz} = \left(\sum_{g,k=1}^{G,K} \frac{N_{gk}}{\psi_{gj}}\left(\mathbf{S}_{gk}\boldsymbol{\beta}_g^{nz\prime} + \left(\bar{\mathbf{x}}_g - \boldsymbol{\tau}_k\right)\boldsymbol{\alpha}_{gk}^{nz\prime}\right)_{j\cdot}\right)\left(\sum_{g,k=1}^{G,K} \frac{N_{gk}}{\psi_{gj}}\left(\boldsymbol{\Theta}_{gk}^{nz} + \boldsymbol{\alpha}_{gk}^{nz}\boldsymbol{\alpha}_{gk}^{nz\prime} + \boldsymbol{\beta}_g^{nz}\left(\bar{\mathbf{x}}_g - \boldsymbol{\mu}_{gk}\right)\boldsymbol{\alpha}_{gk}^{nz\prime}\right)\right)^{-1} \tag{23}$$

$$diag\left(\hat{\boldsymbol{\Psi}}_g\right) = diag\left(\frac{1}{N_g}\sum_{k=1}^{K} N_{gk}\left(\mathbf{S}_{gk} - \left(2\boldsymbol{\Lambda}\boldsymbol{\beta}_g\mathbf{S}_{gk} - \boldsymbol{\Lambda}\boldsymbol{\Theta}_{gk}\boldsymbol{\Lambda}'\right)\right)\right) \tag{24}$$

$$\hat{\boldsymbol{\Phi}}_g = \frac{1}{N_g}\sum_{k=1}^{K} N_{gk}\boldsymbol{\Theta}_{gk} \tag{25}$$

where $\boldsymbol{\Theta}_{gk} = \boldsymbol{\Phi}_g - \boldsymbol{\beta}_g\boldsymbol{\Lambda}\boldsymbol{\Phi}_g + \boldsymbol{\beta}_g\mathbf{S}_{gk}\boldsymbol{\beta}'_g$. Note that the factor loadings (Equation 23) are updated per variable *j* and that the '*nz*' superscript refers to the fact that, for MMG-CFA, the update only concerns the non-zero (nz) loadings for variable *j* – whereas the other loadings, obviously, remain zero. For MMG-EFA, all loadings are non-zero for all variables *j* = 1, …, *J*. When defining a binary 'design' matrix **B** of size *J* × *Q*, with ones for the non-zero loading positions, $\boldsymbol{\beta}_g^{nz}$ is equal

to the columns of $\boldsymbol{\beta}_g$ corresponding to the ones in row j of matrix $\mathbf{B}$ (i.e., $\mathbf{b}_j$), $\boldsymbol{\alpha}_{gk}^{nz}$ contains the elements of $\boldsymbol{\alpha}_{gk}$ corresponding to the ones in $\mathbf{b}_j$, and $\boldsymbol{\Theta}_{gk}^{nz}$ is obtained by selecting rows and columns of $\boldsymbol{\Theta}_{gk}$ as indicated by the ones in $\mathbf{b}_j$.

To set the scale of the factors, the mean factor variances are fixed to one over all groups, i.e., $diag\left(\boldsymbol{\Phi}\right) = diag\left(\frac{1}{N}\sum_{g=1}^{G} N_g \boldsymbol{\Phi}_g\right) = \mathbf{1}$. Note that, for MMG-EFA, the factor covariances are fixed to zero over all groups, which implies that the initial rotation is orthogonal. Afterwards, one can choose to rotate the exploratory factors according to an orthogonal or oblique rotation criterion and counterrotate the group-specific factor covariances accordingly. Additionally, the factor means are restricted per cluster as follows: $\frac{1}{N_k}\sum_{g=1}^{G} N_g \hat{z}_{gk} \boldsymbol{\alpha}_{gk} = \mathbf{0}$ for k = 1, …, K. To impose these restrictions, the factor (co)variances, loadings, factor means and intercepts are transformed as follows:

$$\hat{\boldsymbol{\Phi}}_g^{*} = \left(\boldsymbol{\Theta}\right)^{-\frac{1}{2}} \hat{\boldsymbol{\Phi}}_g \left(\boldsymbol{\Theta}\right)^{-\frac{1}{2}} \quad with \quad \boldsymbol{\Theta} = \frac{1}{N}\left(\sum_{g'=1}^{G}\sum_{k=1}^{K} N_{g'k} \boldsymbol{\Theta}_{g'k}\right) \tag{26}$$

$$\hat{\boldsymbol{\Lambda}}^{*} = \hat{\boldsymbol{\Lambda}}\left(\boldsymbol{\Theta}\right)^{\frac{1}{2}} . \tag{27}$$

$$\hat{\boldsymbol{\alpha}}_{gk}^{*} = \left(\boldsymbol{\Theta}\right)^{-\frac{1}{2}} \hat{\boldsymbol{\alpha}}_{gk} . \tag{28}$$

$$\hat{\boldsymbol{\alpha}}_{gk}^{**} = \hat{\boldsymbol{\alpha}}_{gk}^{*} - \hat{\boldsymbol{\alpha}}_k \quad with \quad \hat{\boldsymbol{\alpha}}_k = \frac{1}{N_k}\sum_{g'=1}^{G} N_{g'} \hat{z}_{g'k} \hat{\boldsymbol{\alpha}}_{g'k}^{*} \tag{29}$$

$$\hat{\boldsymbol{\tau}}_k^{*} = \left(\sum_{g=1}^{G} N_{gk} \boldsymbol{\Sigma}_g^{-1}\right)^{-1} \left(\sum_{g=1}^{G} N_{gk} \boldsymbol{\Sigma}_g^{-1} \left(\bar{\mathbf{x}}_g - \hat{\boldsymbol{\Lambda}}^{*} \hat{\boldsymbol{\alpha}}_{gk}^{**}\right)\right) \tag{30}$$

where the offdiagonal elements of $\boldsymbol{\Theta}$ are fixed to zero in case of MMG-CFA. Note that this is done in the final iteration only (i.e., upon convergence).

A1. Algorithm

For a user-specified number of starts, perform the following steps for each start:

1. Start from a pre-selected random partition (Section A2), i.e., with binary values for $\hat{z}_{gk}$.

2. For each cluster *k* of the random partition, initialize $\boldsymbol{\tau}_k$ as follows: $\boldsymbol{\tau}_k = \frac{1}{N_k}\sum_{g=1}^{G}\hat{z}_{gk}\sum_{n_g=1}^{N_g}\mathbf{x}_{n_g}$.

3. Initialize $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}_g$ and $\boldsymbol{\Phi}_g$, based on probabilistic principal component analysis (Tipping & Bishop, 1999), as follows[6]: $\boldsymbol{\Lambda} = \mathbf{U}_Q\sqrt{\left(\mathbf{V}_Q - \hat{\sigma}^2\mathbf{I}_Q\right)}$ and $\boldsymbol{\Psi}_g = \hat{\sigma}^2\mathbf{I}_J$ for g = 1, …, G, where the columns of $\mathbf{U}_Q$ correspond to the first Q eigenvectors and the diagonal matrix $\mathbf{V}_Q$ contains the first Q eigenvalues of the eigenvalue decomposition of $\mathbf{S} = \frac{1}{N}\sum_{k=1}^{K}\sum_{g=1}^{G}\hat{z}_{gk}\left(\sum_{n_g=1}^{N_g}\tilde{\mathbf{x}}_{n_gk}\tilde{\mathbf{x}}'_{n_gk}\right)$ with $\tilde{\mathbf{x}}_{n_gk} = \mathbf{x}_{n_g} - \boldsymbol{\tau}_k$, $\hat{\sigma}^2$ is the average variance in the $J - Q$ discarded dimensions, and $\mathbf{I}_Q$ and $\mathbf{I}_J$ are $Q \times Q$ and $J \times J$ identity matrices, respectively. In case of MMG-CFA, the thus obtained loadings $\boldsymbol{\Lambda}$ are orthogonally Procrustes rotated towards the design matrix **B** and then the (assumed-to-be) zero loadings are fixed to zero. The factor (co)variance matrices $\boldsymbol{\Phi}_g$ are initialized as $\mathbf{I}_Q$.

4. Initialize $\boldsymbol{\alpha}_{gk}$ for g = 1, …, G and k = 1, …, K based on the singular value decomposition $\tilde{\mathbf{X}}_{gk}\boldsymbol{\Lambda} = \mathbf{U}_{gk}\mathbf{D}_{gk}\mathbf{V}'_{gk}$ where $\tilde{\mathbf{X}}_{gk}$ contains the group-specific deviations around $\boldsymbol{\tau}_k$ and $\mathbf{U}_{gk}$ and $\mathbf{V}_{gk}$ contain the left and right singular vectors, respectively: Group- and cluster-specific component scores are obtained by $\mathbf{F}_{gk} = \left(\mathbf{U}_{gk}\right)_Q\mathbf{V}'_{gk}$ (ten Berge, 1993) where

[6] These starting values are similar to the maximum likelihood estimates of image factor analysis described by Jöreskog (1969).

$\left(\mathbf{U}_{gk}\right)_Q$ corresponds to the first Q columns of $\mathbf{U}_{gk}$. After rescaling $\mathbf{F}_{gk}$ to have a variance of one, $\boldsymbol{\alpha}_{gk}$ is initialized as the mean of $\mathbf{F}_{gk}$.

5. Iterate the following steps while δ_1 & $\delta_2 > 1\times10^{-4}$ and $v \leq 100$:
    a. Update the iteration number: $v = v+1$.
    b. Cycle 1, E-step: Update the posterior classification probabilities $\hat{z}_{gk}$ (Equation 11) for $g = 1, \ldots, G$ and $k = 1, \ldots, K$.
    c. Cycle 1, conditional M-steps:
        i. Update $\hat{\pi}_k$ (Equation 13) for $k = 1, \ldots, K$.
        ii. Update $\hat{\boldsymbol{\tau}}_k$ (Equation 14) for $k = 1, \ldots, K$.
        iii. Update $\hat{\boldsymbol{\alpha}}_{gk}$ (Equation 15) for $g = 1, \ldots, G$ and $k = 1, \ldots, K$.
    d. Cycle 2: E-step: Update the posterior classification probabilities $\hat{z}_{gk}$ (Equation 11) for $g = 1, \ldots, G$ and $k = 1, \ldots, K$.
    e. Cycle 2: Conditional M-steps:
        i. Update $\hat{\pi}_k$ (Equation 20) for $k = 1, \ldots, K$.
        ii. Update $\hat{\boldsymbol{\tau}}_k$ (Equation 21) for $k = 1, \ldots, K$.
        iii. Update $\hat{\boldsymbol{\alpha}}_{gk}$ (Equation 22) for $g = 1, \ldots, G$ and $k = 1, \ldots, K$.
        iv. Update the factor loadings $\boldsymbol{\Lambda}$ (Equation 23), the unique variances $\boldsymbol{\Psi}_g$ for $g = 1, \ldots, G$ (Equation 24) and the factor (co)variance matrices $\boldsymbol{\Phi}_g$ for $g = 1, \ldots, G$ (Equation 25). To remedy near-Heywood cases, fix unique variances to .0001 when they are smaller than this number.
    f. Compute the value of log L^v (Equation 4).

g. Evaluate convergence with respect to log L^v and the parameter estimates $\hat{\theta}_p^v$:

$$\delta_1 = \sum_{p=1}^{P} \left| \frac{\hat{\theta}_p^v - \hat{\theta}_p^{v-1}}{\hat{\theta}_p^{v-1}} \right| \text{ and } \delta_2 = \log L^v - \log L^{v-1}.$$

6. After (preliminary) convergence is reached (or 100 iterations), check whether the obtained solution is the best one in terms of log L (across all starts up to now) and, if so, save the parameter estimates $\hat{\theta}_p^{best} = \hat{\theta}_p^v$ and iteration number $v^{best} = v$.

After performing this procedure for all starts, iterate further until full convergence is reached for the best solution: i.e., starting from $\hat{\theta}_p^v = \hat{\theta}_p^{best}$ and $v = v^{best}$, iterate Steps 5a to 5g while δ_1 & $\delta_2 > 1\times10^{-6}$ and $v \leq 1000$ (or another user-specified maximal number of iterations). Finally, the model identification constraints are imposed by means of Equations 26 to 30.

A2. Multistart procedure

Because the AECM algorithm described in Section A1 is not guaranteed to converge to the global maximum, a multistart procedure is used to increase the probability of finding the global maximum. The multistart procedure applies a tiered testing strategy with respect to several sets of starting values. Specifically, given the user-specified number of starts (e.g., 10), it starts from 10 times as many random partitions of the groups (e.g., 10 × 10 partitions). For each of these partitions, the parameter estimates are initialized as described in Steps 2 to 4 of the algorithm (Section A1). Subsequently, the parameter estimates are updated once by means of Equations 20 to 25 and the log L value (Equation 4) is determined. The 10% most promising partitions (i.e., with the highest log L) are selected as the starts for the algorithm described in Section A1.

Appendix B: Latent Gold 6.0 syntax

An example syntax for MMG-FA with three clusters and two factors for a data set with 11 variables is given and explained below (for more details, see Vermunt & Magidson, 2013):

```
options
   algorithm
      tolerance=1e-006 emtolerance=1e-006 emiterations=1000 nriterations=0
      emfa;
   startvalues
      seed=0 sets=100 tolerance=1e-004 iterations=50 PCA annealing;
   bayes
      categorical=0 variances=0 latent=0 poisson=0;
   missing includeall;
   output
      iterationdetail parameters=first standarderrors probmeans=posterior
      reorderclasses;
variables
   groupid Country;
   dependent (V1-V11) continuous;
   independent Country nominal;
   latent
      F1 continuous,
      F2 continuous,
      Cluster nominal group 4; // 1-8 to estimate models with 1 to 8 clusters
equations
// group-specific factor (co)variances
   F1 | Country;
   F2 | Country;
   F1 <-> F2 | Country;
// logistic regression model for clusters (only intercept)
   Cluster <- 1;
// regression models for factors: group- and cluster-specific factor means
   F1 <- 1 | Country Cluster;
   F2 <- 1 | Country Cluster;
// regression models for items: cluster-specific intercepts for non-referent
// items and invariant loadings
   V1 <- (1) F1;
   V2 - V6 <- 1 | Cluster + F1;
   V7 <- (1) F2;
   V8 - V11 <- 1 | Cluster + F2;
// group-specific unique variances
   V1 - V11 | Country;
```

The LG syntax contains three sections, i.e., ‘options, ‘variables’, and ‘equations’. Firstly, the ‘options’ section pertains to specifications regarding the estimation process and to output options. The parameters in the ‘algorithm’ subsection indicate when the algorithm should proceed with Newton-Raphson instead of EM iterations and when convergence is reached. To apply only

EM iterations, set ‘nriterations’ to zero, ‘emiterations’ to a high number and ‘emtolerance’ to the same value as ‘tolerance’. The option ‘emfa’ makes sure that the factor model parameters are estimated by means of the time-efficient EM procedure detailed in Appendix A. The ‘startvalues’ subsection includes the parameters pertaining to the multistart procedure used by LG. Specifically, for each set of starting values (the number of sets is specified by ‘sets’), the model is re-estimated for as many iterations as specified by ‘iterations’ or until δ_1 or δ_2 drops below the ‘tolerance’ value. Subsequently, it continues with the 10% (rounded upwards) most promising sets (i.e., with the highest log *L*), performing another two times the specified number of iterations (i.e., 2 × the value of ‘iterations’). Finally, it continues with the best solution until convergence. In the example syntax above, 100 starts are requested by the user. The thus obtained multistart procedure is actually more elaborate than the one evaluated in Section 3, because the clustering and factor parameters are updated 50 times before choosing the best 10% most promising sets of starting values (as opposed to one update of the factor parameters only). ‘PCA’ prompts LG to use starting values for the factor loadings and unique variances that are based on a principal component analysis (PCA) of the data (Vermunt and Magidson, 2016). When group sizes are large, the algorithm may be prone to local maxima because the posterior classification probabilities quickly approach one and zero, even for a clustering that is far from the one that is actually underlying the data. This may happen especially when between-cluster intercept differences are small. To avoid this, the ‘annealing’ option – referring to ‘deterministic annealing’ – is used which implies that an auxiliary variable is used to keep the posterior classification probabilities more fuzzy for the first few iterations (Zhou & Lange, 2010). It is advised to set the options in the ‘bayes’ subsection to zero when using the ‘emfa’ algorithm. In the ‘output’ subsection, the user can specify the desired output.

Secondly, the ‘variables’ section specifies the different types of variables included in the model. Since MMG-FA operates on multigroup data, after ‘groupid’, the variable in the data file that indicates the group structure (i.e., the group number for each observation) should be specified, using its label in the data file (e.g, ‘Country’). In the ‘dependent’ subsection, the dependent variables of the model (i.e., the observed variables) are specified, by means of their label in the data file and their measurement scale. Next, the ‘independent’ variables are listed. For MMG-FA, one has to include the grouping variable as an independent variable, since some parameters vary across groups. Finally, the ‘latent’ variables of the MMG-FA model are the factors (i.e., ‘F1’ to ‘F2’ in the example syntax) and the mixture model clustering (i.e., ‘Cluster’). In particular, the former are specified as continuous latent variables, whereas the latter is specified as a nominal latent variable at the group level with a specified number of categories (i.e., the desired number of clusters). For estimating models with, for instance, one to eight clusters, use ‘1–8’.

In the ‘equations’ section, the model equations are listed. First, the factor variances and covariances are specified and they are allowed to differ among groups by adding ‘| Country’. Next, a logistic regression model for the categorical latent variable ‘Cluster’ is specified followed by regression equations for the factors, which contain only an intercept term and where the factor means are made group- and cluster-specific by adding ‘| Country Cluster’ to the factors’ intercept term. Then, regression models are defined for the observed variables, indicating which variables are regressed on which factors. Note that, to apply EFA, all variables should be regressed on all factors. In Latent Gold 6.0, rotation options are available to specify how the parameters should be identified in case of EFA (De Roover & Vermunt, 2019). To obtain intercepts that differ between clusters, ‘| Cluster’ is added to the intercept term. Note that, in the example syntax, V1 and V7 serve as referent items with intercepts of zero in all clusters (i.e., no intercept term included in their

regression equations) and factor loadings of one. This model identification strategy differs from the one in Appendix A. Finally, unique variances are added, which are allowed to differ across groups. At the end of the syntax, additional restrictions may be specified or starting values for all parameters may be given, either by directly typing them in the syntax or by referring to a text file.

Appendix C: Conservation and Self-Transcendence items PVQ-21

Conservation items (male version):

- ipfrule (item 7, conformity): He believes that people should do what they're told. He thinks people should follow rules at all times, even when no one is watching.
- ipbhprp (item 16, conformity): It is important to him always to behave properly. He wants to avoid doing anything people would say is wrong.
- ipmodst (item 9, tradition): It is important to him to be humble and modest. He tries not to draw attention to himself.
- imptrad (item 20, tradition): Tradition is important to him. He tries to follow the customs handed down by his religion or his family.
- impsafe (item 5, security): It is important to him to live in secure surroundings. He avoids anything that might endanger his safety.
- ipstrgv (item 14, security): It is important to him that the government ensures his safety against all threats. He wants the state to be strong so it can defend its citizens.

Self-transcendence items (male version):

- iphlppl (item 12, benevolence): It's very important to him to help the people around him. He wants to care for their well-being.
- iplylfr (item 18, benevolence): It is important to him to be loyal to his friends. He wants to devote himself to people close to him.
- ipeqopt (item 3, universalism): He thinks it is important that every person in the world should be treated equally. He believes everyone should have equal opportunities in life.
- ipudrst (item 8, universalism): It is important to him to listen to people who are different from him. Even when he disagrees with them, he still wants to understand them.
- impenv (item 19, universalism): He strongly believes that people should care for nature. Looking after the environment is important to him.

References to the appendices

De Roover, K., & Vermunt, J. K. (2019). On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-19.

Jöreskog, K. G. (1969). Efficient estimation in image factor analysis. *Psychometrika*, *34*, 51-75.

McLachlan, G., & Krishnan, T. (2007). *The EM algorithm and extensions (Vol. 382).* John Wiley & Sons.

Meng, X. L., & Van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*, 511-567.

ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis.* Leiden, the Netherlands: DSWO Press.

Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal components. *Journal of the Royal Statistical Society B*, *61*, 611 – 622.

Vermunt, J. K., & Magidson, J. (2013). *Technical Guide for Latent GOLD 5.0: Basic, Advanced, and Syntax.* Belmont, MA: Statistical Innovations Inc.

Zhou, H., & Lange, K. L. (2010). On the bumpy road to the dominant mode. *Scandinavian Journal of Statistics*, *37*(4), 612-631.