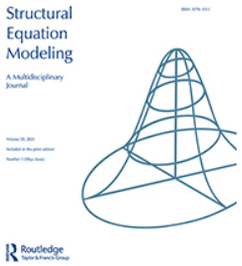
Routledge
Taylor & Francis Group



Structural Equation Modeling: A Multidisciplinary Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hsem20

Testing for Factor Loading Differences in Mixture Simultaneous Factor Analysis: A Monte Carlo Simulation-Based Perspective

Elena Geminiani, Eva Ceulemans & Kim De Roover

To cite this article: Elena Geminiani, Eva Ceulemans & Kim De Roover (2021) Testing for Factor Loading Differences in Mixture Simultaneous Factor Analysis: A Monte Carlo Simulation-Based Perspective, Structural Equation Modeling: A Multidisciplinary Journal, 28:3, 391-409, DOI: 10.1080/10705511.2020.1807351

To link to this article: https://doi.org/10.1080/10705511.2020.1807351


View supplementary material


Published online: 18 Nov 2020.


Submit your article to this journal


Article views: 199


View related articles


View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=hsem20

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2021, VOL. 28, NO. 3, 391–409
https://doi.org/10.1080/10705511.2020.1807351

Routledge
Taylor & Francis Group

Check for updates

Testing for Factor Loading Differences in Mixture Simultaneous Factor Analysis: A Monte Carlo Simulation-Based Perspective

Elena Geminiani [1], Eva Ceulemans [2], and Kim De Roover [2,3]

[1]University of Bologna; [2]KU Leuven; [3]Tilburg University

ABSTRACT

Factor analysis is ubiquitously applied in behavioral sciences for capturing covariances of observed variables by latent variables (factors). When factor-analyzing data from many groups of subjects, mixture simultaneous factor analysis (MSFA) determines which groups have the same factor model by clustering them based on their factor loadings, factor (co)variances and residual variances. Two Monte Carlo simulations are performed to investigate the power and type I error of Wald tests for factor loading differences in MSFA, as affected by characteristics of the data (sample size in terms of number and size of groups), factor models (item communality levels, sizes and types of loading differences) and clustering (cluster size, classification error and uncertainty). The results were better in case of equal cluster sizes, strongly overdetermined factors, high communalities, and larger primary loading differences.

KEYWORDS

Factor analysis; hypothesis testing; mixture clustering; multiple testing

Introduction

In the last decades, factor analysis (Lawley & Maxwell, 1962) has been extensively applied, e.g., in the behavioral sciences, to summarize the interrelationships among the observed variables by means of a smaller set of latent variables, called common factors. Factor analysis can be used in a confirmatory way (CFA, Jöreskog, 1967) by testing an a priori assumed factor structure or in an exploratory way (EFA, Mulaik, 2009) by letting the factors emerge freely as a result of the analysis.

Several kinds of data have a multigroup or multilevel structure (Muthén, 1994; Snijders & Bosker, 1999), e.g., questionnaire items for inhabitants of different countries (Kuppens et al., 2006), or test scores for pupils from different schools (Ma & Klinger, 2000). In these examples, the countries and schools constitute the higher-level units, which are also referred to as "groups."

Many interesting research questions pertain to between-group dissimilarities in the factor models. For instance, in personality psychology, an ongoing debate about the configuration of personality measures opposes the theories of "Big Five" (Goldberg, 1990) versus "Big Three" (Eysenck & Eysenck, 1975). Relatedly, one may wonder whether the personality structure of one culture is the same as that of another (De Raad et al., 2010), that is, whether the same Big Five personality traits are found in every country or whether some countries are characterized by a distinct structure, like the Big Three. Similarly, many cross-cultural psychologists argue that the structure of emotions may differ between cultures (Eid & Diener, 2001; Rodriguez & Church, 2003).

Traditionally, multigroup factor analysis (MGFA, Dolan et al., 2009) is employed to establish the equivalence of factor loadings (i.e., the linear relations between variables and factors) across groups and can be applied in a confirmatory (Jöreskog, 1971) or exploratory way (Hessen et al., 2006). In this MGFA framework, hypothesis testing is often performed by comparing two nested models: the completely *restricted* model, where all factor loadings are restricted to be the same across the groups and the completely *unrestricted* model, where all factor loadings are allowed to differ. If factor loadings equality is not confirmed, and the groups are expected to share a similar structure with one or a few variables behaving differently (Bauer, 2017; De Roover, Timmerman et al., 2017; De Roover et al., 2014), loading-specific testing (Cheung & Rensvold, 1999) is conducted to find the sources of heterogeneity. However, if a specific loading is found to be variant across groups, it is hard to determine for which groups it is actually different and how. Of course, pairwise comparisons of factor loadings may be performed, but in case of a large number of groups they turn out to be cumbersome and not insightful (Kim et al., 2017). Also, the risk of capitalization on chance, due to multiple testing problems, is high (Rutkowski & Svetina, 2014), whereas correcting for this huge multiple testing problem suppresses the power.

A sensible choice to model the unobserved heterogeneity between the groups is then a mixture approach, which assigns the groups to a few "clusters" according to their factor model. Within each cluster, the structures are similar, whereas structures of different clusters are dissimilar. This is accomplished by mixture simultaneous factor analysis (MSFA, De Roover, Vermunt et al., 2017), a multilevel latent variable model with latent variables at both the observation level (the common factors) and the group level (the mixture clustering of groups). A unique feature of MSFA is that it models between-group differences in the covariance structure within each group irrespective of between-group differences in the means (i.e., the

CONTACT Elena Geminiani geminianielena@gmail.com; elena.geminiani4@unibo.it Department of Statistical Sciences, University of Bologna, Bologna 40126, Italy.
Supplemental data for this article can be accessed on the publisher's website.
© 2020 Taylor & Francis Group, LLC

latter do not affect the clustering). This covariance structure heterogeneity (Ansari et al., 2002) is related but not identical to measurement invariance (Meredith, 1993), or rather the lack thereof. Specifically, MSFA clusters groups based on their factor loadings, factor (co)variances and residual or "unique" variances of the variables, whereas differences in factor (co) variances are not violations of measurement invariance.

Conveniently, the stochastic framework of MSFA allows a Wald test for each loading. Whether these tests correctly identify factor loading differences depends on the recovery of the cluster-specific loadings on the one hand, and on the power to detect differences in these loadings on the other hand. In the factor analysis literature, several studies evaluated the influence of a variety of conditions on the quality of factor solutions (see e.g., Hogarty et al., 2005; MacCallum et al., 1999; Pennell, 1968). Factor recovery and precision of loading estimates are impacted by the sample size, which in the current context is made up of both the number of groups and the group size (i.e., the number of subjects within a group). This effect, however, generally reduces in presence of high levels of item communalities (i.e., the proportion of variance in the variables accounted for by the factors) and factor overdetermination (i.e., the variables-to-factors ratio). With respect to likelihood ratio tests of factor loading equivalence in multigroup and mixture CFA, Meade and Bauer (2007) and Buzick (2010) reported that the power for detecting heterogeneous factor loadings depends, among other things, on the size of the true loading differences (i.e., the effect size), and the pattern of loading differences. Within the MSFA framework, we expect the Wald tests to be impacted by the same effects, as well as by other characteristics related to the latent class nature of MSFA, e.g., cluster size and classification error and uncertainty.

This paper evaluates the performance of Wald test statistics to detect specific factor loading differences across clusters of groups in MSFA. To assess how the above effects play out for MSFA, we conducted two extensive Monte Carlo simulation studies and evaluated the "empirical power" (whether true differences were detected) and type I error (false positives). Our study differs from past research on the topic in the following aspects. First, it deals with unknown clusters of groups, whereas existing studies consider either known groups (as in multigroup factor analysis; see Meade & Bauer, 2007) or unknown clusters of subjects, and hence, "unknown groups" (as in mixture CFA; see Buzick, 2010). By clustering several groups (and thus increasing the "within-cluster" sample size), the Wald tests within the MSFA framework are expected to have more power than for multigroup or mixture CFA approaches. Another novel aspect lies in the way hypothesis testing is conducted, that is, through loading-specific testing as opposed to overall tests for all factor loadings at once. The use of loading-specific testing requires multiple testing corrections which, in turn, lower the power. Specifically, we considered two different multiple testing corrections, that is, Bonferroni (1936) and Benjamini and Hochberg (1995) adjustments, and assessed the resulting power results at two significance levels. Finally, because MSFA applies exploratory rather than confirmatory factor analysis within the clusters, we can evaluate the power for detecting all kinds of loading differences, including differences in the cross-loadings and the pattern of primary loadings, unlike existing CFA-based studies (Asparouhov & Muthén, 2014; Muthén & Asparouhov, 2013). Using exploratory factor analysis may, however, imply more error fitting, which may affect both the power for finding these differences and type I error. We focus on the Wald test since we are interested in the mentioned effects on power and type I error, and not in comparing different hypothesis tests (e.g., likelihood ratio test, score test) in this respect.

The remainder of the paper is organized as follows. First, we review MSFA and demonstrate it through an application from emotional psychology. Secondly, we discuss the topic of hypothesis testing for factor loading differences and give more details on the conditions expected to have an impact on their power. Next, Wald tests are illustrated for loading differences in the empirical example, raising relevant questions. We then present two simulation studies on the performance of Wald test statistics for detecting loading differences. We conclude the paper with directions for further research.

Mixture simultaneous factor analysis

We first discuss the type of data mixture simultaneous factor analysis (MSFA) is applicable to and their pre-processing. Then, we describe the characteristics of MSFA and introduce an empirical application.

Data structure and pre-processing

MSFA deals with multivariate multilevel data. In the multigroup case, *Level 1* pertains to the subjects and *Level 2* is constituted by the groups. Both lower-level and higher-level units are assumed to be a random sample of the population of lower- and higher-level units, respectively. Formally speaking, the data of I groups are collected in as many $N_i \times J$ data matrices $\mathbf{X}_i$ $(i = 1, \ldots, I)$ containing the scores on J variables. The number of observations N_i within a group $(n_i = 1, \ldots, N_i)$ may differ across groups, whereas $N = \sum_{i=1}^{I} N_i$ is the overall number of observations. The data of the I groups are concatenated into a global $N \times J$ data matrix $\mathbf{X}$.

MSFA aims to model between-group differences and similarities in the covariance structure within each group (the *within* structure, Timmerman, 2006), irrespective of differences and similarities in the mean levels of the variables per group (the *between* structure). Therefore, MSFA requires that the data are centered per group[1] such that the item means within each group are equal to zero.

Model specification and estimation

MSFA assumes the groups to be sampled from a mixture of K multivariate normal distributions with covariance matrices $\mathbf{\Sigma}_k$ $(k = 1, \ldots, K)$. As already mentioned, the uniqueness of

[1]An alternative would be to include group-specific intercepts. This does not affect the obtained solution.

MSFA over existing multilevel latent variable models lies in the fact that the clustering is solely based on within-group covariances. To this end, cluster-specific means are restricted to be zero, which corresponds to the fact that each group is column-wise centered. Formally, we denote with f the total population density function and with $\boldsymbol{\theta}$ the whole set of parameters. Similarly, f_k refers to the kth cluster-specific density function and $\boldsymbol{\theta}_k$ to its parameters. The mixing proportions, i.e., the prior probabilities of belonging to cluster k, are indicated by π_k, which are subject to a unit sum constraint, $\sum_{k=1}^{K} \pi_k = 1$. At the group level, the MSFA model is formulated as follows:

$$f(\boldsymbol{X};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{X}_i;\boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k \prod_{n_i=1}^{N_i} \mathcal{N}_J(\mathbf{0};\boldsymbol{\Sigma}_k), \qquad (1)$$

where $\mathcal{N}_J$ is the J-variate normal density function. For each cluster, a factor model is estimated. In particular, for the n_i^{th} observation belonging to cluster k, the J-dimensional vector of scores $\boldsymbol{x}_{n_i|k}$ is decomposed as follows:

$$\boldsymbol{x}_{n_i|k} = \boldsymbol{\Lambda}_k \boldsymbol{\eta}_{n_i} + \boldsymbol{e}_{n_i}, \qquad (2)$$

where $\boldsymbol{\Lambda}_k$ refers to the cluster-specific matrix of factor loadings, $\boldsymbol{\eta}_{n_i}$ denotes the Q_k-dimensional vector of factor scores (with Q_k the number of common factors of the kth cluster) and $\boldsymbol{e}_{n_i}$ the J-dimensional vector of residuals. Factor loadings quantify how much each factor affects the scores on each variable. Factor scores and residuals are both assumed to be multivariate normally distributed, that is, $\boldsymbol{\eta}_{n_i} \sim \mathcal{N}_{Q_k}(\mathbf{0}, \boldsymbol{\Psi}_k)$ and $\mathbf{e}_{n_i} \sim \mathcal{N}_J(\mathbf{0}, \mathbf{D}_k)$, where $\boldsymbol{\Psi}_k$ is the $Q_k \times Q_k$ factor covariance matrix for cluster k, and $\mathbf{D}_k$ is the $J \times J$ diagonal matrix containing the unique variances (i.e., variable-specific variation including measurement error) of the variables in cluster k. Each cluster-specific factor model implies that $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Psi}_k \boldsymbol{\Lambda}'_k + \mathbf{D}_k$. For the sake of simplicity, De Roover, Vermunt et al. (2017) considered an equal number of factors across clusters (i.e., $Q_k = Q$ for $k = 1, \ldots, K$). Note that imposing an equal number of factors makes perfect sense when looking for a few loading-specific differences among sets of factors. Also, in order to capture all differences in covariance structure by means of the clustering and cluster-specific loadings (and unique variances), orthogonality of the factors was imposed for each group and thus for each cluster (i.e., $\boldsymbol{\Psi}_i = I_q$ for $i = 1, \ldots, \mathbf{I}$ and thus $\boldsymbol{\Psi}_k = \mathbf{I}$ for $k = 1, \ldots, K$, where $\mathbf{I}_q$ is the $Q \times Q$ identity matrix). In case of orthogonal factors, factor loadings can be interpreted as covariances between the observed variables and the factors, and the cluster-specific covariance matrices become $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k \boldsymbol{\Lambda}'_k + \mathbf{D}_k$. The path diagram of an MSFA model with two orthogonal factors per cluster is depicted in Figure 1.

Even though MSFA allows to trace differences in factor model that are substantively interesting or that may cause incorrect inferences when enforcing one overall factor model to the data (Ansari et al., 2002), it is less suited to find measurement non-invariances because differences in factor (co)variances are also captured by the clustering while these are irrelevant to measurement invariance. Each set of cluster-specific factors has rotational freedom which we removed in the simulations and the empirical application by imposing $q - 1$ zero loadings for factor $q = 1, \ldots, Q$ (Echelon form, Dolan et al., 2009). Alternatively, rotational criteria (Kaiser, 1958; Kiers, 1997; De Roover & Vermunt, 2019) can be used to enhance the interpretability and/or comparability of the loadings.

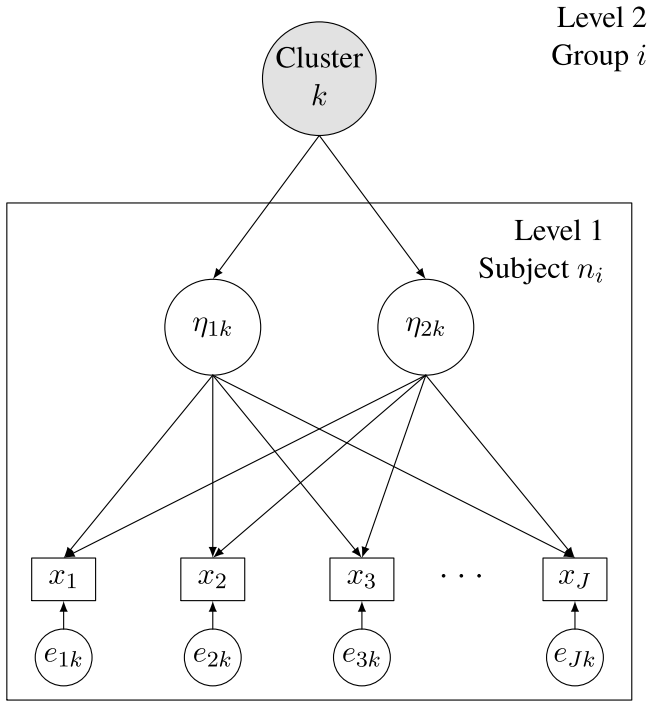


Figure 1. Path diagram of an MSFA model with two orthogonal factors per cluster. The nested frames reflect the multilevel nature of MSFA. Following the conventions, the empty circles represent the continuous latent variables, the filled circle denotes a categorical latent variable, and the rectangles represent the observed variables. The arrows connecting the factors and the observed variables represent linear relations. Intercepts are omitted from the diagram since the data are centered per group. Due to the exploratory nature of MSFA, every factor is connected to every observed variable. When using oblique rotation, cluster-specific factor correlations are also present.

For details on the maximum likelihood estimation for MSFA, the reader is referred to De Roover, Vermunt et al. (2017). Conveniently, an MSFA model can be estimated in Latent GOLD (Vermunt & Magidson, 2013) by means of a proper syntax file. In the Appendix, the Latent GOLD syntaxes used for the simulation studies and the empirical application are presented.

Empirical example

To illustrate the research questions regarding factor loading differences in MSFA, we recap the application in De Roover, Vermunt et al. (2017). The data set stems from the 2001 International College Survey (ICS; Kuppens et al., 2006) and concerns 9-point Likert scale ratings of emotions of positive (*contentment, happy, cheerful, pride, gratitude, love*) and negative (*sad, anger, guilt, shame, worry, stress, jealousy*) valence in terms of their normativeness. Specifically, the question was "*How appropriate and valued is each of the following emotions in your society? Do people approve of this emotion?*". The sample included 8894 participants from 48 nations. The number of subjects within each country ranged from 26 to 531.

Table 1. Clustering of the countries of the MSFA model with two clusters and two factors per cluster for the Emotion Norm data from the 2001 ICS study. When present, classification uncertainties are given in brackets.

Cluster 1	Bangladesh, Brazil, Bulgaria, Cameroon, Georgia, Ghana, India, Iran, Nepal, Nigeria, Slovakia (0.9980), South Africa (0.9965), Thailand, Turkey, Uganda, Zimbabwe
Cluster 2	Australia, Austria, Belgium, Canada, Chile, China, Colombia, Croatia, Cyprus, Egypt, Germany, Greece, Hong Kong, Hungary, Indonesia, Italy, Japan, Kuwait, Malaysia, Mexico, Netherlands, Philippines, Poland, Portugal, Russia, Singapore, Slovenia, South Korea, Spain, Switzerland, United States, Venezuela

An MSFA model with two clusters and two orthogonal factors per cluster was fitted to the data, where the secondary loading of *jealousy* was restricted in both clusters to zero to eliminate rotational freedom. The best-fitting partition allocated 16 countries (composed of 2015 subjects) to the first cluster, and 32 (composed of 6879 subjects) to the second one (Table 1), which led to very unequal cluster sizes (approximately 23% and 77% of the subjects, respectively). The classification probabilities were always either zero or one, except for Slovakia and South Africa, which had (very negligible) uncertainties of 0.9980 and 0.9965, respectively. De Roover and colleagues reported that Cluster 1 pertained to the developing conservative countries and Cluster 2 to the developed progressive countries.

Table 2 reports the estimated cluster-specific factor loadings for the two clusters, the absolute differences among the corresponding loadings, as well as the cluster-specific unique variances. In general, the two loading structures are very similar. The extent to which a factor is similar in the two factor loading matrices can be measured through Tucker's congruence coefficient (Tucker, 1951). The congruences were equal to 0.98 for both factors, which implies that they are nearly identical, considering that Lorenzo-Seva and Ten Berge (2006) report that two factors are highly similar when the congruence coefficient is at least 0.95. The negative emotions load on the first factor and the positive ones on the second factor. Thus, it seems that the first factor corresponds to disapproved emotions, whereas the second factor represents the emotions that are considered appropriate. *Pride*, however, has a very different loading pattern for the two clusters: its primary loading is on Factor 1 in the first cluster, and on Factor 2 in the second cluster. This suggests that pride is considered a socially disapproved emotion for the countries in the first cluster, i.e., as bragging. For the countries in Cluster 2 it is a sign of self-confidence, which is judged as appropriate. Moreover, the factor loadings are in absolute value mostly higher in Cluster 1 than in Cluster 2. In addition to visual inspection, Wald tests are useful to determine which loading differences are significant.

Wald tests for factor loading differences in MSFA

In this paper, we apply the Wald test to test the equality of factor loadings across clusters in MSFA. In this context, the Wald test compares the model with all loadings free to vary across clusters (in this case, except for the zero-restricted loading to achieve Echelon rotation) to the model with one loading restricted to be invariant. In agreement to the empirical application, we consider two clusters ($K = 2$). In this case,[2] the null hypothesis can be expressed as:

$$\mathrm{H}_0 : \lambda_{jq}^1 = \lambda_{jq}^2 \qquad \text{for} \quad j = 1, \ldots, J, \quad q = 1, 2, \tag{3}$$

where λ_{jq}^1 and λ_{jq}^2 denote the factor loadings of the jth item on the qth factor in the first and second cluster, respectively. Under the alternative hypothesis, the examined factor loading differs across clusters, that is, $\mathrm{H}_1 : \lambda_{jq}^1 \neq \lambda_{jq}^2$. The Wald test statistic (Asparouhov, 2007; Wald, 1943) can then be written as:

$$W_{jq} = \frac{\left(\hat{\lambda}_{jq}^1 - \hat{\lambda}_{jq}^2\right)^2}{\mathrm{Var}\left(\hat{\lambda}_{jq}^1 - \hat{\lambda}_{jq}^2\right)} \qquad \text{for} \quad j = 1, \ldots, J, \quad q = 1, 2, \tag{4}$$

Table 2. Estimated unique variances and cluster-specific factor loading matrices of the MSFA model with two clusters and two factors per cluster and the corresponding absolute differences for the Emotion Norm data from the 2001 ICS study.

	Unique variances		Factor Loadings				Absolute differences	
			Cluster 1		Cluster 2			
Items	Cluster 1	Cluster 2	Negative	Positive	Negative	Positive	Negative	Positive
Contentment	1.47	3.48	−0.25	**1.44**	−0.13	**1.21**	0.12	0.23 *†
Happy	0.63	1.39	−0.26	**1.60**	−0.17	**1.42**	0.09	0.18 *†
Love	1.21	2.37	−0.26	**1.39**	−0.08	**1.22**	0.18	0.17 †
Sad	2.76	4.19	**1.32**	−0.32	**1.26**	0.07	0.06	0.39 *†
Jealousy	2.85	4.94	**1.29**	*0*	**1.27**	*0*	0.02	–
Cheerful	1.53	2.38	−0.30	**1.18**	−0.06	**1.04**	0.24 †	0.14 †
Worry	2.01	2.86	**1.74**	−0.07	**1.43**	0.07	0.31 *†	0.14
Stress	2.15	2.63	**2.01**	−0.25	**1.69**	−0.16	0.32 *†	0.09
Anger	1.87	2.23	**1.97**	−0.37	**1.54**	−0.16	0.43 *†	0.21
Pride	3.41	5.33	**1.10**	0.27	0.34	**0.61**	0.76 *†	0.34 *†
Guilt	2.80	4.42	**1.24**	0.05	**1.10**	0.13	0.14	0.08
Shame	3.01	4.85	**1.03**	0.18	**1.07**	0.10	0.04	0.08
Gratitude	2.88	3.95	−0.29	**0.95**	−0.14	**0.86**	0.15	0.09

Note: The items are ratings on a 9-point Likert scale to the question *"How appropriate and valued is each of the following emotions in your society? Do people approve of this emotion?"*. For each emotion, the primary loading is shown in bold, whereas the restricted loadings are in italic and underlined. Absolute differences with asterisks (*) and daggers (†) indicate that the corresponding factor loading difference is significant at 1% level after Bonferroni and Benjamini-Hochberg correction, respectively.

[2]In the general case with K clusters, there is a Wald statistic for each cluster pair and the null hypothesis corresponds to all of them being equal to zero.

where $\hat{\lambda}^1_{jq} - \hat{\lambda}^2_{jq}$ is the maximum likelihood estimator (MLE) of $\lambda^1_{jq} - \lambda^2_{jq}$, and $\text{Var}\left(\hat{\lambda}^1_{jq} - \hat{\lambda}^2_{jq}\right)$ represents its variance. The MLE $\hat{\lambda}^1_{jq} - \hat{\lambda}^2_{jq}$ is asymptotically normal with mean $\lambda^1_{jq} - \lambda^2_{jq}$ (i.e., it is asymptotically unbiased) and variance $\frac{1}{\mathcal{I}(\lambda^1_{jq} - \lambda^2_{jq})}$ (i.e., it is asymptotically efficient), where $\mathcal{I}(\lambda^1_{jq} - \lambda^2_{jq})$ is the Fisher (or "expected") information matrix. A consistent – and preferred (Buse, 1982; Efron & Hinkley, 1978) – estimator of this variance is given by the inverse of the observed information matrix evaluated at the MLE point, i.e., $\frac{1}{\mathcal{J}(\hat{\lambda}^1_{jq} - \hat{\lambda}^2_{jq})}$, which is the one used in this paper. The asymptotic normal distribution of the MLE implies an asymptotic chi-squared distribution with 1 degree of freedom (in general, $K - 1$ degrees of freedom) for the Wald statistics under the null hypothesis $W|\text{H}_0 \underset{N\to\infty}{\longrightarrow} \chi^2_1$. The p-value is then the right-tailed chi-squared probability above the observed value for the Wald test statistic.

Power and type I error

Rejection of H_0 when it is true is called a type I error (occurring with probability α), whereas acceptance of H_0 when it is false is a type II error (occurring with probability β). The complement of β (i.e., $1 - \beta$) is the power of the test, i.e., the probability that the test correctly rejects the null hypothesis when the alternative is true. The specificity is the complement to one of the type I error rate α. Ideally, a good test should have high values of both power and specificity, but in practice every test is subject to a trade-off between type I and type II errors. In this paper, we focus on the empirical power, which refers to the proportion of simulated data sets under the same design conditions (i.e., over multiple replicates) for which the null hypothesis is correctly rejected, and evaluate it under a broad range of conditions. Following the convention generally adopted in the literature (Cohen, 1988), we consider 80% as the desired level of power. In the remainder of this section, we first discuss the multiple testing problem and corrections that may be applied, followed by a number of factors that may affect the power and type I error of the Wald test for factor loading differences in MSFA.

Multiple testing problem

Since a Wald statistic is computed for each factor loading difference, expression (3) results in a set of $M = JQ$ (minus $\frac{Q(Q-1)}{2}$, in case of zero-restricted loadings) null hypotheses being tested simultaneously. When multiple tests are performed at the same time, the probability of a type I error increases. A simple solution consists in applying a Bonferroni correction, which tests each individual hypothesis at a $\frac{\alpha}{M}$ significance level. For example, in case of a 1% level, all the hypotheses with a p-value $\leq \frac{0.01}{M}$ are rejected. By doing so, the probability of falsely rejecting at least one true null hypothesis (*familywise error rate*, FWER) is guaranteed not to exceed α. This causes a loss in the power for detecting differences, however. An alternative is to control the *false discovery rate* (FDR[3]), i.e., the expected proportion of falsely rejected null hypotheses among those that are rejected. The Benjamini-Hochberg (BH) method ensures that the FDR does not exceed α. Specifically, let $p_{(1)} \leq \ldots \leq p_{(h)} \leq \ldots \leq p_{(M)}$, with $h = 1, \ldots, M$, denote the p-values ranked in increasing order, and $H_0^{(1)}, \ldots, H_0^{(h)}, \ldots, H_0^{(M)}$ the corresponding null hypotheses. For each p-value, the critical value $\alpha\frac{h}{M}$ is computed. Then, the BH rejection threshold is found as the largest p-value that is smaller than its critical value, that is, the method seeks that p-value $p_{(L)}$ for which $L = \max\left\{h : p_{(h)} \leq \alpha\frac{h}{M}\right\}$. All the null hypotheses whose p-values are below the BH rejection threshold are rejected. For the empirical application (with $\text{M} = 13 * 2 - 1$) and $\alpha = 0.01$, the BH method, starting from the smallest p-value, finds the last p-value falling below the line $0.01\frac{h}{25}$, for $h = 1, \ldots, 25$, and rejects all the hypotheses with smaller or equal p-values. Due to its less stringent control on error rates, the BH method always shows greater power – at the expense of higher type I error rates – than the Bonferroni correction. Despite this necessary multiple testing correction, the Wald tests for factor loading differences in MSFA still have more power and lower type I error rates than the same tests in MGFA. The power is increased because the clustering of the groups augments the sample size for each factor model, whereas the decrease in type I error occurs since less factor loadings are compared, thus leaving less room for possible idiosyncratic differences due to error fitting. In this respect, the unavoidable decrease in power due to multiple testing corrections is much less marked in MSFA than in MGFA, where a possibly huge number of pairwise comparisons is needed.

Total sample size

Based on findings from empirical studies in factor analysis literature (Browne & Cudeck, 1992; Hogarty et al., 2005), the condition expected to have the strongest effect on power is the sample size, which in MSFA is made up by both the number of groups and the group size. The sample size is also important for the asymptotic normality of the Wald test (under H_0). Due to the multilevel nature of the data (i.e., multiple observations per group), the total sample quickly becomes large and the Wald test might, at some point, become "overpowered" and indicate even the smallest differences in factor loadings as statistically significant (Chatfield, 1995), thus making the test more prone to type I errors. In the application, there were 8894 subjects divided in 48 countries with a mean group size of approximately 185 subjects.

Within-cluster sample size

For MSFA, the power of the Wald tests is also affected by the within-cluster sample sizes, which depend on the number of clusters and the clustering itself. Specifically, the power may depend strongly on the size(s) of the smallest cluster(s) (Gudicha et al., 2016). In the application, we observed very unequal cluster sizes, with nearly 23% of the subjects allocated to the first cluster and 77% to the second one.

[3]FDR $\leq$ FWER, thus any approach controlling the FWER, like the Bonferroni correction, also controls FDR.

Level of the communality of the variables

This is the proportion of variance in the variable accounted for by the factors. In general, higher levels of communality are associated with a higher power (Hogarty et al., 2005; MacCallum et al., 1999; Pennell, 1968). Thus, for the application, we expect the power of Wald tests to be higher for the loading differences of, e.g., *stress* and *anger* (high communality) in contrast to those of, e.g., *cheerful* and *gratitude* (low communality).

Factor overdetermination

This is the degree to which the factor is clearly represented by a sufficient number of variables. Overdetermined factors have high loadings on at least three to four variables and exhibit good simple structure (Hogarty et al., 2005) and imply a higher power for detecting loading differences than less determined factors. For equal overdetermination levels, this effect is reduced as communalities increase, and may also interact with sample size (MacCallum et al., 1999). Considering the application, this implies that the power of Wald tests is expected to be higher for the differences in Factor 1 (more overdetermined) than for the ones in Factor 2.

Size of factor loading difference

With respect to likelihood ratio tests of factor loading equivalence in multigroup and mixture CFA, Meade and Bauer (2007) and Buzick (2010) report that the power for detecting factor loading differences also depends on the size of the true loading differences, i.e., on the effect size. We expect to find the same effect for Wald tests in case of MSFA. For the application, this translates to the fact that the difference between the primary loadings of *pride* (i.e., 0.76) will be much more easily detected than the difference between those of *cheerful* (i.e., 0.14).

Type of factor loadings

Not only the size of the loading difference, but also whether it concerns primary or secondary loadings can affect the power. On the one hand, Pennell (1968) showed that the power for detecting a loading difference increases when the comparison loading is higher. On the other hand, the presence of a cross-loading for a certain variable enlarges the standard error for the primary loading of that variable and thus decreases the power for detecting differences in the latter. Thus, the power for detecting the 0.18 difference between the cluster-specific primary loadings of *happy* is expected to be higher than the power for detecting the equally sized difference between the secondary loadings of *love*.

Cluster separation

In absence of cluster-specific intercepts and factor means, as in MSFA, heterogeneous factor loadings are more difficult to detect, due to the lower separability of the clusters (Buzick, 2010). On top of that, less clusters, equal cluster sizes, and a larger number of items increase cluster separation and thus the power (Gudicha et al., 2016). In the application, the cluster-separation may be lowered by the markedly unequal cluster sizes, among other things.

Classification uncertainty and misclassification

As a first aspect related to cluster separation, the power of Wald tests in MSFA or other latent class models also depends on the amount of uncertainty about the unobserved subjects' cluster memberships. In presence of classification uncertainty, the estimated observed information $\mathcal{J}(\hat{\lambda}^1_{jq} - \hat{\lambda}^2_{jq})$ becomes smaller, thus resulting in a lower power. Note that De Roover and colleagues report very little classification uncertainty for MSFA. Another aspect related to the cluster separation is misclassification. In MSFA, when the clustering of the groups is not recovered correctly, neither will the factor loadings, lowering power to detect loading differences. Classification uncertainty was minimal for the application, whereas misclassification was unknown, as for any real data example.

The effects of the above aspects and their interactions are evaluated in the simulations below.

Wald tests for the empirical example

For the empirical application, Wald tests were performed to discover which of the loadings are statistically different across the clusters (Table 2). Absolute differences with asterisks (*) and daggers (†) indicate that the corresponding loading difference is statistically significant at 1% level after Bonferroni and BH correction, respectively. As expected, the BH correction allows to detect more differences (i.e., the difference between the primary loadings of *love* and the two differences for *cheerful*) than the Bonferroni correction. For both factors, in addition to the remarkable aforementioned differences for *pride*, smaller differences are found to be significant as well. The majority of the detected differences are between primary loadings, which are always higher in the first cluster. This indicates that *worry, stress* and *anger* and *contentment* and *happy* have more common variance to be explained by the negative and positive factors, respectively, in developing conservative countries than in the developed progressive societies. For the secondary loadings, we only find a significant difference for *sad* (and also for *cheerful* with BH correction), where the stronger and negative cross-loading in the first cluster suggests that in developing countries this negative emotion is inversely related to positive emotions. Note that, although an equal number of differences is detected for the two factors, more subtle differences are identified for the second one: for instance, differences of 0.23 and 0.18 (for *contentment* and *happy*, respectively) are significant for Factor 2 at 1% level after Bonferroni correction, whereas differences of 0.24 and 0.18 (for *cheerful* and *love*, respectively) are not significant for Factor 1. This may be due to the level of the communalities being higher for *contentment* and *happy* (1.81 and 2.34, respectively) than for *cheerful* and *love* (1.28 and 1.75, respectively), and to the type of loading difference, namely, between primary loadings instead of cross-loadings. In this case, the higher overdetermination of the first factor was probably not sufficient to identify the factor loading differences in question. All absolute differences above 0.30 are found to be significant at 1%

level after Bonferroni correction, whereas all differences below 0.14 are not. In addition, only some of the absolute differences between 0.14 and 0.30 are found to be significant.

Considering these Wald test results, two main research questions arise:

(1) Research question 1: To what extent do the power and type I error of the Wald tests depend on the within-cluster sample sizes, as determined by the number of groups, group sizes, and the way observations are divided across clusters?
(2) Research question 2: To what extent do the power and type I error of the Wald tests depend on characteristics of the loading structures being compared? To what extent are they affected by the difference in factor over-determination between the two factors? To what extent are they affected by differences in item communality? And to what extent are they affected by the nature of the difference itself (primary or cross-loading difference)?

Method

An effective way of evaluating the performance of the Wald tests for the detection of factor loading differences across MSFA clusters is through Monte Carlo simulations. We conducted two Monte Carlo simulations to examine how the performance is affected by the number of groups, the group size, the level of item communalities, the absolute size of factor loading differences and the type of loading it concerns, as well as by effects that follow from the latent class nature of MSFA – i.e., cluster sizes, classification errors, and classification uncertainty. The first simulation study addresses research question 1 specifically for the empirical factor loadings in Table 2, whereas the second simulation study addresses both research questions 1 and 2 and how the two groups of aspects interact in a more controlled setting. Notably, the first simulation study is the more empirically valid one, whereas the second one makes use of more artificial loading patterns yet allows to disentangle the effects mentioned in research question 2 from one another. Additionally, in simulation study 2, only a subset of the factor loadings is different across the clusters, making it possible to also evaluate the type I errors of the Wald tests in a range of conditions, whereas in simulation study 1 all loadings differ between the clusters implying that only power can be evaluated.

Monte Carlo simulation

Monte Carlo studies (Bandalos & Gagné, 2012) allow to evaluate the Wald test performance under a broad range of potentially influential factors by generating sample data matrices from models with known population parameters. They require the specification of the so-called Monte Carlo variables (characteristics of the data) and population variables (characteristics of the model; see below). Common Monte Carlo variables include the sample size and the number of replications. The total sample size in MSFA was manipulated by systematically varying the following conditions:

- Cluster sizes: groups are either equally (50%-50%) or unequally (20%-80%) divided across clusters. The unequal cluster sizes nearly match those of the empirical example, with Cluster 1 (with the highest loadings in general) always being the smallest cluster;
- Number of groups: 10, 30, and 50, where 50 corresponds closely to the 48 countries in the empirical data set;
- Number of observations within each group: 100, 200, and 400. These within-group sample sizes correspond closely to those investigated by MacCallum et al. (1999) and Meade and Bauer (2007) and 200 is close to the mean group size of 185 in the empirical data set.

The first simulation study constitutes a 2 (cluster size) $\times$ 3 (number of groups) $\times$ 3 (group size) factorial design, resulting in 18 conditions. The second simulation additionally investigates the effects of two population variables: the level of item communalities, and the absolute size of factor loading differences and the type of loading it concerns. Therefore, the second study constitutes a 2 (cluster size) $\times$ 3 (number of groups) $\times$ 3 (group size) $\times$ 2 (item communalities) $\times$ 3 (difference pattern) design, resulting in 108 conditions. For each cell of the two simulation designs, 100 replicates of the data were generated, so that a total of 1800 data matrices were obtained for the first study, and 10,800 data matrices for the second one.

Data generation models

Both simulations are inspired by the empirical application therefore, the number of variables ($J = 13$), clusters ($K = 2$) and factors ($Q = 2$) exactly match those of the data analysis. In the first study, the cluster-specific loading matrices and the unique variances coincide with the ones in Table 2. Specifically, for *pride*, the primary loading shifts from Factor 1 to Factor 2 causing loading differences of 0.76 and 0.34 on Factor 1 and Factor 2, respectively, whereas other differences in primary and cross-loadings range from 0.02 to 0.43. This method of aligning the population parameters to the estimates resulting from fitting a model to an actual data set has been endorsed by MacCallum (2003). While the advantage of this technique is that it more likely reflects real-world conditions, and therefore produces results generalizable to those conditions, a limitation lies in the impossibility of manipulating all the model characteristics of interest. For this reason, the second simulation consists of an experimental study, where all variables of interest can be easily manipulated. These variables include:

- Level of item communalities: factor loading differences were induced for either high- or low-communality items;
- Type of difference: primary loading decrease, weak cross-loading, and strong cross-loading (explained below). The primary loading decrease corresponds to the "uniform" pattern applied in Buzick (2010).

Specifically, for simulation study 2, the population parameters of Cluster 1 were generated as follows. In the loading matrix of

Cluster 1, the items either loaded exclusively on the first or on the second factor with primary loadings initially set equal to one and secondary loadings equal to zero. In order to be consistent with the data analysis, eight items loaded on the first factor, and the remaining ones on the second one. The first 7 variables were assigned a *high* communality of 0.8, whereas the last 6 items a *low* communality of 0.4. This corresponded to roughly 40% of error and to a mean communality of 0.615 for each generated data set. Next, the factor loadings were scaled

$$\boldsymbol{\Lambda}_1^T = \begin{array}{c} \\ \text{Factor1} \\ \text{Factor2} \end{array} \begin{array}{c} \begin{array}{cccccccccccc} V_2 & V_3 & V_4 & V_5 & V_6 & V_7 & V_8 & V_9 & V_{10} & V_{11} & V_{12} & V_{13} & V_{14} \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0.89 & 0.89 & 0.89 & 0.89 & 0.63 & 0.63 & 0.63 & 0.63 & 0 & 0 \\ 0.89 & 0.89 & 0.89 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.63 & 0.63 \end{bmatrix} \end{array},$$

according to these item-specific communalities, producing the following factor loading matrix of Cluster 1:
where the numbering of the items started from the second variable since the first one, V_1, is the nominal indicator of the group membership in the Latent GOLD syntax. Note that these are strongly overdetermined factors according to the definition by Hogarty et al. (2005). Due to their better simple structure, they are more overdetermined in comparison to the empirical loadings used in the first simulation. The unique variances of Cluster 1 were chosen such that each variable had a population-level variance of 1.0, in agreement with the studies of Meade and Bauer (2007) and Buzick (2010). The loadings of Cluster 2 were generated by changing the size of some loadings of Cluster 1. Differences were created for four items, two of which loading on Factor 1 and the other two on Factor 2. The selected items were V_3, V_4, V_6 and V_7 for the high-communality case and V_9, V_{11}, V_{13} and V_{14} for the low-communality case. Regarding the manner in which their loadings were modified in Cluster 2, the types of difference patterns are:

- Primary loading decrease, or "primary": only the primary loadings of all the selected variables are modified by decreasing them by a specific constant (the difference size), which varied per variable (details are given below);
- Weak cross-loading, or "cross-weak": in this pattern both the primary and secondary loadings are modified. The primary loadings of two items (V_4 and V_7 or V_9 and V_{13}, depending on the high- or low-communality case) were strongly decreased, whereas their secondary loadings were slightly increased, creating a small cross-loading. The size of the introduced difference varied per item and per loading (details are given below);
- Strong cross-loading, or "cross-strong": similarly, in this condition, the primary loadings of two items (V_4 and V_7 or V_9 and V_{13}) were slightly decreased, whereas their secondary loadings were strongly increased, creating a moderate cross-loading (details are given below).

As mentioned above, several difference sizes were explored (0.1, 0.2, 0.3 and 0.4) within each of the difference patterns. In order to obtain the same average difference (0.25) for both factors, the sizes of the created difference were 0.2 and 0.3 for the items loading on the first factor, and 0.1 and 0.4 for the items loading on the second factor. However, we should acknowledge that, as item communalities correspond to the sum of the squared factor loadings for each variable, a certain change for a higher factor loading constituted a larger change in communality than did an equal change in a lower factor loading (Meade & Bauer, 2007). The factor loading matrices for Cluster 2, $\boldsymbol{\Lambda}_2$, resulting from the aforementioned manipulations of the loadings of the selected high- and low-communality items are shown in Table 3 for each difference pattern. It is important to note that the factor loading differences induced by these patterns cannot be traced by CFA-based methods, which do not allow for different cross-loadings and patterns of primary loadings across groups.

Data analysis models

Latent GOLD ® (version 5.1) was called to simulate each data matrix $\mathbf{X}$, employing the cluster-specific population parameters as starting values. Due to sampling fluctuations, the generated data were not perfectly centered per group. Therefore, each data matrix was pre-processed in R (R Core Team, 2020) in order to columnwise center all its groups. It was then analyzed in Latent GOLD by estimating an MSFA model with the correct number of clusters ($K = 2$) and factors ($Q = 2$). In the first simulation, the secondary loading of *jealousy* was restricted to zero in both clusters to remove rotational freedom. In the second simulation, since the manipulated loading matrices showed a simple structure, many loadings were zero in both clusters and hence could be restricted to zero; for convenience, the secondary loading of V_{14} was chosen to be restricted. The algorithm was run with 25 random starts and the best solution was retained. According to the model specifications discussed above, the cluster-specific factors were estimated to be uncorrelated and with unit variance for each group. The Latent GOLD syntaxes used for generating the data and fitting the MSFA models are given in the Appendix.

Evaluation measures

To examine the *recovery of the cluster membership of groups*, we compute the adjusted Rand index (ARI, Hubert & Arabie, 1985) between the true and the estimated[4] (modal) clustering of the groups. The ARI is a measure of cluster agreement

[4]Each group is assigned to the cluster for which the posterior probability is the highest.

Table 3. The factor loading matrices of Cluster 2 in simulation study 2 by communality level and difference pattern. The manipulated loadings are shown in bold, whereas the restricted loadings are in italic and underlined.

	High communality						Low communality					
	Primary		Cross-Weak		Cross-Strong		Primary		Cross-Weak		Cross-Strong	
Items	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2	F1	F2
V_2	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89
V_3	0	**0.79**	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89
V_4	0	**0.49**	**0.20**	**0.49**	**0.30**	**0.79**	0	0.89	0	0.89	0	0.89
V_5	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0
V_6	**0.69**	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0
V_7	**0.59**	0	**0.59**	**0.10**	**0.69**	**0.40**	0.89	0	0.89	0	0.89	0
V_8	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0	0.89	0
V_9	0.63	0	0.63	0	0.63	0	**0.43**	0	**0.33**	**0.10**	**0.43**	**0.40**
V_{10}	0.63	0	0.63	0	0.63	0	0.63	0	0.63	0	0.63	0
V_{11}	0.63	0	0.63	0	0.63	0	**0.33**	0	0.63	0	0.63	0
V_{12}	0.63	0	0.63	0	0.63	0	0.63	0	0.63	0	0.63	0
V_{13}	0	0.63	0	0.63	0	0.63	0	**0.53**	**0.20**	**0.23**	**0.30**	**0.53**
V_{14}	*0*	0.63	*0*	0.63	*0*	0.63	*0*	**0.23**	*0*	0.63	*0*	0.63

corrected for chance and is one when there is perfect agreement between the two clusterings and zero when agreement is at chance level.

The *recovery of cluster loading matrices* is assessed via the goodness-of-cluster-loading-recovery statistic (GOCL). This index is obtained by computing Tucker's congruence coefficients between the corresponding factors of the true and estimated loading matrices and averaging across factors and clusters as follows:

$$GOCL = \frac{\sum_{k=1}^{K}\sum_{q=1}^{Q}\phi(\Lambda_{k(q)}, \hat{\Lambda}_{k(q)})}{KQ}, \quad (5)$$

where $\Lambda_{k(q)}$ and $\hat{\Lambda}_{k(q)}$ refer, respectively, to the qth factor of the true and estimated factor loading matrices of cluster k. The GOCL statistic takes values between 0 (no recovery at all) and 1 (perfect recovery). The GOCL statistic can also be applied to assess the recovery of a specific factor, that is, $GOCL_q = \frac{\sum_{k=1}^{K}\phi(\Lambda_{k(q)}, \hat{\Lambda}_{k(q)})}{K}$, for $q = 1, \ldots, Q$. In the second simulation, the between-cluster congruence coefficients in the population for Factor 1 are equal to 0.99 in all difference patterns. For Factor 2, the coefficients are equal to 0.98 and 0.97 (depending on high- or low-communality level, respectively) in the primary and cross-weak patterns, and to 0.97 in the cross-strong pattern. These congruences pertain to a high similarity between clusters (Lorenzo-Seva & Ten Berge, 2006) and correspond nicely to the congruences in the empirical analysis.

As previously discussed, *classification uncertainties* negatively affect the power of a hypothesis test for between-cluster differences, and therefore they are monitored by tracking the lowest posterior classification probability per group, and then computing the maximum of these uncertainties for each data set.

The *empirical power* of Wald tests is assessed by computing the proportion of data sets across the simulated conditions for which a specific loading difference is significant at a certain significance level after a given multiple testing correction. For Bonferroni correction at 1% level, each hypothesis is rejected at the adjusted 1% level if its corresponding p-value does not exceed $\frac{0.01}{13*2-1} = 0.0004$ (note that the number of tests equals the number of loadings minus one, due to the zero-loading restriction). Contrarily, with the BH correction we would reject the ("ordered") null hypotheses up to $H_0^{(L)}$, where $L = \max\{h : p_{(h)} \leq \alpha\frac{h}{13*2-1}\}$, with the false discovery rate α equal to 0.01. Whereas the power quantifies the proportion of true differences that are correctly identified as such, the *type I error* of the test is calculated as the frequency of loading equalities incorrectly identified as a difference.

To determine the most important effects, analyses of variance (ANOVAs) are conducted for the power and type I error rates across all 100 replications of each cell of the design. To investigate which main and interaction effects have a large effect size, we calculate the eta-squared η^2 indices (Cohen, 1988; Olejnik & Algina, 2000) that indicate the proportion of variance in the performance measure that is accounted for by all main, interaction and error effects. Only large effects (i.e., with an eta-squared of at least 0.14, according to the guidelines in Miles & Shevlin, 2001) are discussed.

Results

Goodness of recovery

Simulation study 1

The ARI always equaled one, so the MSFA algorithm perfectly recovered the clustering of the groups. The mean GOCL of 0.998 ($SD = 0.003$) indicated that the recovery of Factor 1 ($Mean = 0.999$, $SD = 0.002$) and Factor 2 ($Mean = 0.998$, $SD = 0.005$) was generally excellent (Lorenzo-Seva & Ten Berge, 2006). Classification uncertainty was not an issue, since the maximum uncertainty ($Mean = 0.000$, $SD = 0.000$) was very low for all data sets.

Simulation study 2

With an overall mean ARI of 0.993 ($SD = 0.072$), the MSFA algorithm clearly recovered the clustering of the groups very well. In particular, for 10,658 data sets (98.7% of the total), the ARI was exactly equal to 1.0. The few data sets for which there was not a perfect agreement between the true and estimated clustering were generally characterized by the smallest number of groups and group size (i.e., 10 groups of 100 observations each), unequal cluster sizes (i.e., 20%-80% condition), low-communality items, and a strong cross-loading difference

pattern. The reason for the latter is that the strong cross-loading condition has the smallest between-cluster difference in unique variance for the selected items, which contributes to a lower separability of clusters. The mean GOCL of 0.999 ($SD = 0.004$) indicated that the recovery of Factor 1 ($Mean = 1.000$, $SD = 0.002$) and Factor 2 ($Mean = 0.998$, $SD = 0.005$) was generally excellent. For just six data sets, the GOCL values were less than 0.95, all of which still corresponded to a high similarity to the true loading matrices. Generally, the lower values of this statistic occurred for data sets with 10 groups, group sizes of 100 observations, loading differences pertaining to low-communality items, and the primary loading difference pattern. This is explained by the facts that firstly, a smaller sample size, as always, worsens the loadings recovery, secondly, low-communality items also have higher unique variances (thus, more error), and thirdly, the cross-loading difference patterns implied a higher level of overdetermination of both factors in comparison to the primary case (i.e., the primary loading manipulations resulted in decreased communalities). Classification uncertainty was not an issue, since the maximum uncertainty ($Mean = 0.000$, $SD = 0.005$) was very low for all data sets.

Power

Simulation study 1

Table 4 presents the empirical power results of the Wald tests for each loading difference at 1% level and after Bonferroni correction. The first column of the table indicates the examined factor loadings, and hence provides information about the type of difference (that is, whether it compares primary loadings or cross-loadings), whereas the second one quantifies the size of the absolute difference. Specifically, the difference sizes are sorted from small to large differences per factor. The third column contains what we will refer to as "mean item communalities", that is, the average across the two clusters of the communalities corresponding to the examined loadings; note that the averaging of the two quantities makes perfect sense, since both communalities have an effect on the power for finding the difference. In order to get a quick overview of the results, the cells with the desired level of power (i.e., ≥ 0.80) are indicated in gray, and the loading differences are split per factor and listed in ascending order by size. Using the BH approach instead of Bonferroni would have resulted in reaching the desired level of power for 23 more differences (8 for Factor 1 and 15 for Factor 2; see Table A.1 in the Supplemental Online Material) in addition to those identified with the Bonferroni correction. These additional differences ranged from 0.14 to 0.43, and had a power after Bonferroni correction between 0.54 and 0.79. Using the 5% instead of the 1% level would have resulted in sufficient power for 18 additional differences with the Bonferroni correction, and for 35 more differences with the BH correction (see Tables A.2 and A.3 in the Supplemental Online Material, respectively).

As expected, as the number of groups and the group size increased, more factor loading differences were detected. The power was also higher when the observations were equally spread into the two clusters (i.e., 50%-50% versus 20%-80% condition). The ANOVAs with power at the 1% level as dependent variable revealed that the number of groups I and the group size N_i had the greatest impact ($\eta^2(I) = 0.54, \eta^2(N_i) = 0.40$ with Bonferroni correction, and $\eta^2(I) = 0.56, \eta^2(N_i) = 0.40$ with BH correction). Clearly, larger absolute difference sizes are associated with a higher power. Specifically, the majority of the large differences (i.e., above 0.30) were detected in case of at least 30 groups of 200 observations, whereas almost none of the small differences (i.e., below 0.14) was found to be significant, even with the largest sample sizes. As far as the medium differences (i.e., ranging from 0.14 to 0.30) are concerned, some were detected (usually with 30 or 50 groups of 400 observations) and others were not. However, many questions remain unanswered; for instance, why did a difference of 0.14 for the second factor and with a low "mean item communality" (1.28) show very good power (0.98) in presence of 50 groups of 400 observations and equal cluster sizes, whereas another difference of the same size, on the same factor, with the same cluster sizes and a higher "mean item communality" (2.54) reached a limited power (0.38)? Is the better power in the former induced by the type of loading difference, i.e., between primary loadings instead of cross-loadings, rather than by the level of "mean item communalities"? If so, this would confirm the findings of Pennell (1968), namely, that the power for detecting a loading difference increases when the comparison loading is higher. Simulation study 2 was conducted to study these effects in a more systematic and untangled manner and answer these questions.

Simulation study 2

Power was inspected in function of the design conditions at the 1% and 5% level (after multiple testing correction), over both factors as well as per factor. The few data sets with an ARI less than 1.0 were evaluated separately, since they were so few in number and we want to disentangle (lack of) power from (lack of) cluster recovery. The total proportion of true positives for these data sets at 1% level was 0.32 after Bonferroni correction, and 0.35 after BH correction. For the data sets with an ARI equal to 1, the power of Wald tests was first examined with respect to the difference sizes and the cluster sizes. Table 5 reports the results with the two multiple testing corrections at 1% level (for the power results at 5% level, see Tables B.1-B.5 in the Supplemental Online Material). Not surprisingly, when the size of the true differences increased, the power became higher. Also, better power results occurred in presence of equal cluster sizes. If a researcher is satisfied with a power above 0.80, then the power for detecting a difference of size 0.1 is insufficient, while for a size of 0.2 or larger the power is sufficient in all conditions and the Wald tests can generally detect almost all true positives at 1% level.

The ANOVAs with power at the 1% level as dependent variable revealed that the number of groups I and the group size N_i had the greatest impact ($\eta^2(I) = 0.47, \eta^2(N_i) = 0.32$ with Bonferroni correction, and $\eta^2(I) = 0.45, \eta^2(N_i) = 0.31$ with BH correction). As I increased, the values of power were higher, and thus the power of Wald tests considerably improved (Table 6). However, this effect appeared to be influential only when the size of the simulated differences was small. In fact, for differences larger than 0.2, power was always (near)

Table 4. From the left, the factor loadings (split per common factor and in ascending order by difference size), the corresponding absolute difference sizes and the mean item communalities in simulation study 1. Then, the mean power by cluster size condition, number of groups, and group size with Bonferroni correction at 1% level. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

			Equal cluster sizes									Unequal cluster sizes								
			10 groups			30 groups			50 groups			10 groups			30 groups			50 groups		
Loadings	Abs. diff.	Mean item comm.	100 obs	200 obs	400 obs	100 obs	200 obs	400 obs	100 obs	200 obs	400 obs	100 obs	200 obs	400 obs	100 obs	200 obs	400 obs	100 obs	200 obs	400 obs
Factor 1																				
1.29/1.27	0.02	1.64	0	0	0	0	0	0.01	0	0	0	0	0	0	0	0	0	0	0	0
1.03/1.07	0.04	1.12	0	0	0	0	0	0.02	0	0	0.03	0	0.01	0	0	0	0	0	0	0
1.32/1.26	0.06	1.72	0	0	0	0	0.01	0.02	0.01	0.01	0.05	0	0	0	0	0.02	0.02	0	0	0.03
−0.26/-0.17	0.09	2.34	0	0	0.01	0.01	0.01	0.01	0	0.02	0.10	0	0	0	0	0.01	0.02	0	0.01	0.02
−0.25/-0.13	0.12	1.81	0	0	0.03	0	0.05	0.06	0.02	0.08	0.21	0	0	0.03	0	0	0.05	0.01	0.01	0.07
1.24/1.10	0.14	1.38	0	0.01	0.03	0.02	0.14	0.53	0.13	0.38	0.84	0	0.01	0.03	0.01	0.05	0.29	0.04	0.23	0.61
−0.29/-0.14	0.15	0.87	0	0	0.01	0.02	0.10	0.24	0.08	0.22	0.67	0	0.03	0.01	0.03	0.01	0.19	0	0.06	0.31
−0.26/-0.08	0.18	1.75	0.01	0.03	0.09	0.03	0.13	0.35	0.11	0.35	0.80	0	0.02	0.07	0.01	0.06	0.29	0.04	0.11	0.54
−0.30/-0.06	0.24	1.28	0.01	0.06	0.36	0.19	0.48	0.97	0.39	0.85	1	0	0.01	0.15	0.09	0.18	0.73	0.18	0.53	0.99
1.74/1.43	0.31	2.54	0.11	0.31	0.92	0.69	1	1	0.96	1	1	0.02	0.29	0.55	0.48	0.92	0.99	0.81	1	1
2.01/1.69	0.32	3.49	0.11	0.33	0.89	0.72	0.99	1	0.96	1	1	0.02	0.14	0.62	0.35	0.83	1	0.64	0.99	1
1.97/1.54	0.43	3.21	0.35	0.78	1	0.98	1	1	1	1	1	0.15	0.54	0.95	0.81	0.99	1	0.95	1	1
1.10/0.34	0.76	0.89	0.86	1	1	1	1	1	1	1	1	0.67	1	1	1	1	1	1	1	1
Factor 2																				
0.18/0.10	0.08	1.12	0	0	0	0	0.01	0	0.01	0.01	0.04	0	0	0	0	0.01	0.04	0.01	0	0.05
0.05/0.13	0.08	1.38	0	0	0.01	0	0	0.03	0	0.02	0.04	0	0	0	0	0	0.02	0.01	0.02	0.02
0.95/0.86	0.09	0.87	0	0	0.02	0	0.02	0.07	0.01	0.04	0.23	0	0	0	0.02	0.01	0.05	0.02	0.04	0.18
−0.25/-0.16	0.09	3.49	0	0	0	0	0	0.01	0	0	0.08	0.01	0	0	0	0.01	0	0	0	0.03
1.18/1.04	0.14	1.28	0	0.05	0.18	0.11	0.32	0.79	0.18	0.57	0.98	0	0.03	0.06	0.05	0.17	0.45	0.06	0.32	0.82
−0.07/0.07	0.14	2.54	0.01	0	0.02	0.01	0.07	0.10	0.02	0.07	0.38	0.01	0	0.01	0.02	0.02	0.03	0.04	0.06	0.15
1.39/1.22	0.17	1.75	0	0.07	0.27	0.21	0.56	0.91	0.40	0.85	1	0	0	0.14	0.11	0.31	0.72	0.17	0.76	0.99
1.60/1.42	0.18	2.34	0.03	0.09	0.41	0.24	0.77	0.97	0.51	0.93	1	0	0.06	0.13	0.10	0.41	0.90	0.34	0.90	1
−0.37/-0.16	0.21	3.21	0.01	0	0.07	0.02	0.09	0.56	0.08	0.29	0.82	0.01	0	0.02	0.04	0.05	0.26	0.05	0.23	0.54
1.44/1.21	0.23	1.81	0.03	0.20	0.61	0.30	0.80	0.99	0.60	0.97	1	0.02	0.11	0.25	0.18	0.70	0.98	0.43	0.96	1
0.27/0.61	0.34	0.89	0.06	0.26	0.73	0.44	0.93	1	0.82	1	1	0.02	0.08	0.37	0.32	0.69	1	0.62	0.97	1
−0.32/0.07	0.39	1.72	0.12	0.30	0.69	0.52	0.93	1	0.85	0.99	1	0.07	0.12	0.44	0.33	0.77	1	0.69	0.99	1

Table 5. Mean power by simulated difference size and cluster size conditions at 1% level after Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction in simulation study 2. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

	Power			
	Equal cluster sizes		Unequal cluster sizes	
Difference size	Bon.	BH	Bon.	BH
0.1	0.52	0.60	0.37	0.46
0.2	0.91	0.93	0.84	0.87
0.3	0.99	0.99	0.96	0.97
0.4	1	1	0.98	0.99

Table 6. Mean power by number of groups and group size in all difference size scenarios with Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction at 1% level in simulation study 2. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

		Difference size							
		0.1		0.2		0.3		0.4	
		Bon.	BH	Bon.	BH	Bon.	BH	Bon.	BH
I	10	0.13	0.20	0.67	0.73	0.92	0.94	0.97	0.98
	30	0.50	0.60	0.96	0.97	1	1	1	1
	50	0.70	0.77	0.99	1	1	1	1	1
N_i	100	0.20	0.28	0.73	0.78	0.93	0.95	0.97	0.98
	200	0.44	0.54	0.90	0.93	0.99	0.99	1	1
	400	0.68	0.75	0.99	0.99	1	1	1	1

1.00, regardless of the number of groups. Similar conclusions can be drawn for the group size. Although the size of their interaction effect was below the cutoff, the power results are given for each combination of difference size, number of groups, and group size (Table 7), as this may be of interest to the reader. The within-cluster sample sizes are added: since there is hardly any classification uncertainty, these values are also informative for exploratory MGFA with these sample sizes within groups.

In contrast to our hypothesis and to findings in previous studies the communality level and difference pattern did not affect the power of the Wald tests very much: the eta-squared was very low for the communality level, whereas, although higher, it did not reach the cutoff value for the difference pattern. As for the difference pattern, all manipulations produced outstanding power results in the presence of big differences (Table 8). Notice that, whereas the primary pattern investigated the power for factor loading differences of 0.1, 0.2, 0.3, and 0.4 between primary loadings, the results for the weak (strong) cross-loading pattern investigated the power for differences of 0.1 and 0.2 between cross-loadings (primary loadings), and of 0.3 and 0.4 between primary loadings (cross-loadings). The weak cross-loading pattern was the one with the worse

Table 7. Mean power by sample and cluster size in all difference size scenarios with Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction at 1% level in simulation study 2. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

			Within-cluster sample size		Power			
			Equal cluster sizes	Unequal cluster sizes	Equal cluster sizes		Unequal cluster sizes	
Difference size	Number of groups	Group size	Cl. 1 – Cl. 2	Cl. 1 – Cl. 2	Bon.	BH	Bon.	BH
0.1	10	100	500 – 500	200 – 800	0.03	0.06	0.00	0.01
		200	1000 – 1000	400 – 1600	0.11	0.21	0.05	0.09
		400	2000 – 2000	800 – 3200	0.39	0.52	0.18	0.30
	30	100	1500 – 1500	600 – 2400	0.25	0.37	0.12	0.19
		200	3000 – 3000	1200 – 4800	0.59	0.71	0.38	0.53
		400	6000 – 6000	2400 – 9600	0.90	0.94	0.72	0.83
	50	100	2500 – 2500	1000 – 4000	0.52	0.63	0.27	0.38
		200	5000 – 5000	2000 – 8000	0.87	0.91	0.66	0.77
		400	10,000 – 10,000	4000 – 16,000	0.98	0.99	0.90	0.94
0.2	10	100	500 – 500	200 – 800	0.42	0.52	0.17	0.25
		200	1000 – 1000	400 – 1600	0.82	0.88	0.60	0.70
		400	2000 – 2000	800 – 3200	1	1	0.94	0.96
	30	100	1500 – 1500	600 – 2400	0.96	0.98	0.80	0.87
		200	3000 – 3000	1200 – 4800	1	1	0.99	0.99
		400	6000 – 6000	2400 – 9600	1	1	1	1
	50	100	2500 – 2500	1000 – 4000	1	1	0.97	0.98
		200	5000 – 5000	2000 – 8000	1	1	1	1
		400	10,000 – 10,000	4000 – 16,000	1	1	1	1
0.3	10	100	500 – 500	200 – 800	0.88	0.92	0.68	0.74
		200	1000 – 1000	400 – 1600	0.99	1	0.94	0.96
		400	2000 – 2000	800 – 3200	1	1	1	1
	30	100	1500 – 1500	600 – 2400	1	1	1	1
		200	3000 – 3000	1200 – 4800	1	1	1	1
		400	6000 – 6000	2400 – 9600	1	1	1	1
	50	100	2500 – 2500	1000 – 4000	1	1	1	1
		200	5000 – 5000	2000 – 8000	1	1	1	1
		400	10,000 – 10,000	4000 – 16,000	1	1	1	1
0.4	10	100	500 – 500	200 – 800	0.96	0.97	0.85	0.88
		200	1000 – 1000	400 – 1600	1	1	0.98	0.99
		400	2000 – 2000	800 – 3200	1	1	1	1
	30	100	1500 – 1500	600 – 2400	1	1	1	1
		200	3000 – 3000	1200 – 4800	1	1	1	1
		400	6000 – 6000	2400 – 9600	1	1	1	1
	50	100	2500 – 2500	1000 – 4000	1	1	1	1
		200	5000 – 5000	2000 – 8000	1	1	1	1
		400	10,000 – 10,000	4000 – 16,000	1	1	1	1

Table 8. Mean power by difference size and difference pattern with Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction at 1% level in simulation study 2. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

	Difference size							
	0.1		0.2		0.3		0.4	
Difference pattern	Bon.	BH	Bon.	BH	Bon.	BH	Bon.	BH
Primary	0.54	0.62	0.91	0.93	0.98	0.99	1	1
Cross-weak	0.27	0.35	0.82	0.85	0.99	0.99	1	1
Cross-strong	0.53	0.61	0.90	0.93	0.95	0.96	0.97	0.98

Table 9. Mean power of the two factors by number of groups and group size with Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction at 1% level in simulation study 2. Cells with the desired level of power (i.e., ≥ 0.80) are colored gray.

		Power			
		Factor 1		Factor 2	
		Bon.	BH	Bon.	BH
I	10	0.79	0.84	0.55	0.59
	30	0.98	0.99	0.75	0.80
	50	1	1	0.85	0.89
N_i	100	0.83	0.86	0.59	0.63
	200	0.95	0.96	0.72	0.77
	400	0.99	1	0.84	0.88

performance for the small difference sizes. Possibly, because the factors have a less simple structure than in the primary condition (Table 3; Pennell, 1968), which decreases the power. Another possible reason pertains to the lower communalities in Cluster 2 than those in the primary and strong cross-loading conditions. The influence of communalities was demonstrated to be stronger for loadings around zero (Pennell, 1968).

Power was also scrutinized separately for each common factor, in order to investigate to which extent the significant effects discovered above applied to each factor, and whether other design conditions turned out to be influential. We noticed that power tended to be higher for the first factor (with Bonferroni correction, $Mean = 0.925$, $SD = 0.264$; with BH correction, $Mean = 0.942$, $SD = 0.234$) than for the second one (with Bonferroni correction, $Mean = 0.717$, $SD = 0.450$; with BH correction, $Mean = 0.759$, $SD = 0.428$). With respect to the first factor, the ANOVAs on its power at 1% level indicated that the design factors that substantially affected the power of the Wald tests for the loadings of the first factor were not only the number of groups and the group size, but also their interaction (with Bonferroni correction, $\eta^2(I) = 0.35$, $\eta^2(N_i) = 0.23$ and $\eta^2(I * N_i) = 0.25$; with BH correction, $\eta^2(I) = 0.31$, $\eta^2(N_i) = 0.21$ and $\eta^2(I * N_i) = 0.27$). The effect of I and N_i and their interaction were the same as the ones previously described for overall power (Table 9). The ANOVAs on the second factor revealed an effect of the number of groups and group size at 1% level (with Bonferroni correction, $\eta^2(I) = 0.44$ and $\eta^2(N_i) = 0.32$; with BH correction, $\eta^2(I) = 0.45$ and $\eta^2(N_i) = 0.32$). The effects of I and N_i were stronger than those for the first factor, but without an interaction effect due to overall sample size.

Type I error

Simulation study 1

No type I error evaluation is possible since all the loadings were different between the two clusters in the empirical application.

Table 10. Frequency table of the number of false positives with Bonferroni (Bon.) and Benjamini-Hochberg (BH) correction at 1% level in simulation study 2 with percentages enclosed in parentheses.

	False Positives					
	0	1	2	3	4	5–12
Bon.	10,581 (99.28)	62 (0.58)	6 (0.06)	4 (0.04)	5 (0.05)	-
BH	10,365 (97.25)	223 (2.09)	34 (0.32)	11 (0.10)	3 (0.03)	22 (0.21)

The maximum possible number of false positives is 21 (26 loadings minus 4 that are different between clusters in the population minus 1 restricted to be zero in both clusters).

Simulation study 2

The frequency distribution of the number of false positives per data set at 1% level is given in Table 10 (for the results at 5% level, see Table B.6 in the Supplemental Online Material). Although the maximum possible number of false positives was 21, the number of false positives was rarely greater than 3. Note that the Bonferroni correction, which presented a lower power, showed excellent results in terms of false positives due to its stringent control on the FWER. None of the investigated conditions was found to have an impact on the number of false positives with Bonferroni correction at 1% level, i.e., none of the eta-squared exceeded the cutoff. The type I error rates of the two factors were also very similar to one another.

Finally, the distributions of the estimated absolute difference sizes were computed for the false positives for the two factors (Figure 2) using Bonferroni correction (BH correction yielded very similar results). The median difference size was 0.11 for Factor 1 and 0.19 for Factor 2. The majority of these estimated differences, for both factors, did not exceed 0.3, suggesting that with a loading difference greater than 0.3, one can be more confident that the detected difference is a true positive.

Summary of the findings

The recovery of the cluster membership of the groups and of the cluster-specific loading matrices was generally excellent.

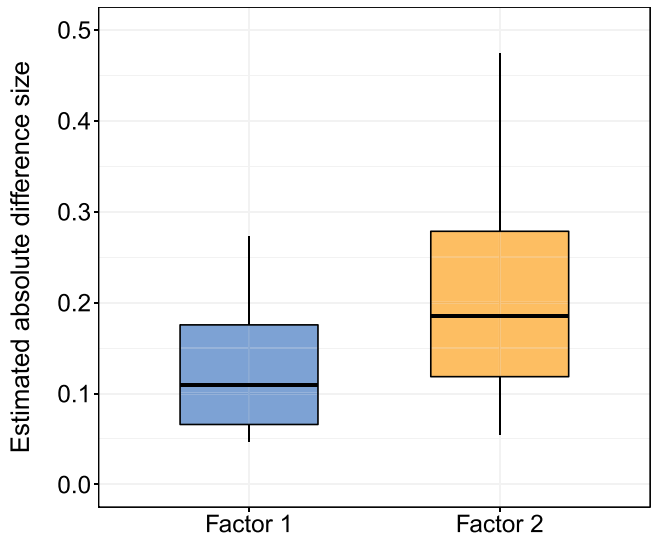


Figure 2. Distributions of the estimated absolute difference sizes of the false positives for the two factors with Bonferroni correction at 1% level in simulation study 2.

Classification uncertainties did not constitute a problem, probably because the sample size per group was sufficient to allocate groups to clusters with certainty. When analyzing data with smaller sample sizes or less separated clusters, researchers may face more classification uncertainties or classification error, which will affect the Wald test performance.

The simulation studies gave much insight on which conditions have to be satisfied so that researchers can be confident that the effects of scientific interest are detected. Specifically, when the effect size, i.e., the size of loading differences, was equal to 0.2 or larger, the Wald tests tended to identify all true positives (no type II errors) for most investigated sample sizes. This is in line with the findings by Meade and Bauer (2007) and Buzick (2010) for multigroup and mixture CFA. To detect smaller differences (i.e., below 0.2), researchers need both a larger number of groups and larger group sizes. In this respect, and as an answer to research question 1, practitioners can look at Table 7 to get an idea about the sample size requirements to detect the difference sizes of interest, or conversely, which differences are likely to be detected with the available sample size. Also, the simulations showed that having equal rather than unequal cluster sizes led to better performances, which replicates the findings by Gudicha et al. (2016). Concerning research question 2, the higher the overdetermination of a factor (i.e., good simple structure with more primary loadings) the higher the power for finding loading differences for that factor, which is in agreement with the findings by Hogarty et al. (2005). Furthermore, a difference between primary loadings was always more easily found than a difference of the same size between cross-loadings, which confirms the findings by Pennell (1968). Item communalities seemingly did not play a role in this respect, probably due to the presence of an already adequate sample size and well-determined factors (MacCallum et al., 1999). It is also interesting to notice that comparable levels of power could be reached under multiple design conditions. This gives researchers the flexibility of designing studies that allow them to detect the difference sizes of interest while taking into account possible limitations in their design settings. For instance, small sample sizes (e.g., due to expenses in recruiting participants) can be compensated for by ensuring a strong overdetermination of the factors and using items with high communalities and complying with a simple structure.

The Wald tests presented an excellent performance, across all manipulated conditions, in correctly retaining the null hypothesis in case of no difference. Thus, type I error rates were not an issue, especially with a lower significance level. This indicates that the Wald tests did not become "overpowered" in presence of the larger sample sizes (the "p-value" problem), at least not within the examined conditions. Moreover, an inspection of the distributions of estimated absolute difference sizes for the false positives revealed that they were generally associated with estimated loading differences not exceeding 0.3. This implies that when a larger effect is found, in conditions similar to the simulated ones, one can be rather confident that the detected difference is a true positive.

The simulations focused on the 1% level, whereas a 5% level produced better power results at the cost of more type I errors (see the Supplemental Online Material). A lower significance level should be preferred when wary of false positives, whereas a higher level when one wants to avoid false negatives. Greene (2003) and Leamer (1978) suggest to decrease the value of significance from usual levels for larger data sets. Of course, the use of multiple testing corrections somewhat reduced the power. In this regard, the Bonferroni method moderately impacted the power in favor of a more stringent control on false positives, whereas the Benjamini-Hochberg procedure produced better power results without losing much in terms of type I errors. The choice of which method to use depends on the scientific goals of investigation: Bonferroni correction is more appropriate when the number of true differences is small (as in simulation 2, where there were 4 truly false null hypotheses) or when the consequence of incorrectly declaring a difference to be significant is grave, whereas Benjamini–Hochberg method is to be preferred when higher levels of power are desired being aware that some mistakes could possibly be made (Foulkes, 2009).

Discussion

Mixture simultaneous factor analysis is a novel method for conveniently exploring covariance structure heterogeneity in terms of differences in factor model across many groups (De Roover, Vermunt et al., 2017). Subsequently, Wald tests can be performed to determine which factor model parameters are significantly different across the obtained clusters of groups. This paper focused on the performance of Wald tests for detecting factor loading differences across MSFA clusters, as these are the parameters that determine how the cluster-specific factors are interpreted. The effect of characteristics of the data (sample size in terms of number of groups and group size), the underlying factor models (item communality levels, sizes and types of loading differences), as well as the latent class nature of MSFA (cluster size, classification error and uncertainty) on the power and type I error of the Wald tests for detecting factor loading differences were mapped. The Wald tests generally showed good power and did not exhibit signs of "overpowering" due to the generally larger sample size implied by the multilevel nature of the data. The Bonferroni correction resulted in a lower power for small difference sizes, yet resulted in a marked improvement in terms of type I error rates. The Benjamini-Hochberg procedure produced a slightly nicer tradeoff between high levels of power and control on the type I errors. In line with the findings from empirical studies (Browne & Cudeck, 1992; Gudicha et al., 2016; Hogarty et al., 2005), the within-cluster sample size, which is determined by the number of groups, the group size, and the cluster sizes, was the most influential factor when it comes to the power for detecting small loading differences. In this respect, the simulation provided guidelines for the necessary sample size in order to obtain sufficient power for detecting small loading differences below 0.2. If the differences of interest are known or expected to be equal to 0.2 or larger, they are likely to be found with sufficient power from the smallest evaluated sample sizes already. On top of that, the distributions of estimated absolute difference sizes for the false positives revealed that they were generally not greater than 0.3 (under the simulated

conditions). The Wald test results thus seem to be most reliable for loading differences of 0.3 or larger.

The occurrence and influence of classification error and uncertainty were minimal in the current simulation studies and will be minimal in comparable empirical data sets. For this reason, the obtained results are also relevant for loading-specific testing in multigroup exploratory factor analysis (i.e., using the cluster sizes as the group sizes). In future research, it would be interesting, however, to investigate the Wald test performance in data sets with less separated clusters and thus more classification error as well as classification uncertainty.

Although the simulation studies exceeded the conditions evaluated in MacCallum et al. (1999), Hogarty et al. (2005), Meade and Bauer (2007), and Buzick (2010) in terms of the variety of the loading structures being compared, they were limited in the range of the levels of the examined settings. Even if it is plausible that the trends found within the tested range behave in a predictable way outside of that tested range, the findings do not formally have external validity beyond the specific explored scenarios (Bandalos & Gagné, 2012). The results may also not generalize to conditions with violations of the model assumptions, for instance, non-normal observed scores or latent variables within clusters (Molenaar & Dolan, 2018).

The simulations were restricted to the scenario with two clusters and two common factors per cluster. Obtaining sufficient power to detect a certain difference across more clusters will require more groups and/or observations per group (i.e., to achieve the same within-cluster sample size and cluster separation). Future research may include MSFA models with more clusters and evaluate Wald testing across pairs of clusters to detect for which clusters loadings are different.

These findings were obtained by estimating an MSFA model with the correct number of clusters and factors. In empirical practice, the most appropriate numbers of clusters and factors need to be selected first. To select the most appropriate number of clusters, popular methods are minimizing the Bayesian Information Criterion (BIC, Schwarz et al., 1978), other information-based indices (e.g., Akaike information criterion, Akaike, 1974; hierarchical BIC, Zhao et al., 2013) or employing the stepwise procedures illustrated in Lukočiene˙ et al. (2010). In presence of non-normal observed scores, spurious clusters could be selected to capture the distributions of the observed variables. In that case, researchers can resort to non-normal mixture models (Muthén & Asparouhov, 2015), which reduce the risk of spurious clusters by relaxing the within-cluster normality assumption (Wickrama et al., 2016). Whereas the effect of extracting too few clusters is clear (i.e., some differences will not be captured and thus cannot be detected with Wald testing), the effect of overextracting clusters is more puzzling. In that case, the recovery and thus also the detection of the factor loading differences depend on how the truly underlying clusters are dispersed across the larger number of retained clusters. In any case, when too many clusters are retained, a decrease in power to detect true differences as well as a decrease in type I error rates is to be expected, because the within-cluster sample sizes are smaller, the cluster separation is lower, and error fitting may mask true differences. Future research needs to focus on model selection for MSFA as well as on how this affects the recovery of the clustering and factor loadings (much like in De Roover et al., 2012). For now, potential users are advised to combine BIC, and the like, with substantive considerations when selecting the number of clusters. With respect to the number of factors, effects of under- and overextraction have been evaluated by Fava and Velicer (1992).

Besides the Wald test, two alternatives exist for determining which factor loading differences are significant in the MSFA framework, that is, the *score test* (Rao, 1948), also known as the *Lagrange multiplier test* (Breusch & Pagan, 1980) or *modification index* (Sörbom, 1989), and the *likelihood ratio test* (Wilks, 1938). This paper focuses on the Wald test because it remains a viable option to evaluate the equality of factor loadings in two extreme conditions; (1) when there is little or no heterogeneity of the loadings across the clusters, a situation that would negatively affect the performance of the likelihood ratio test[5] (Buzick, 2010), and (2) when most of the loadings are heterogeneous over clusters, which would corrupt the score test[6] (Cheung & Rensvold, 1999). Additionally, they are the ones displayed by default in the output of Latent GOLD, and are thus readily available for users. Despite the findings in this discussion specifically pertain to Wald tests, thanks to the asymptotical equivalence to the score and the likelihood ratio tests (Engle, 1984), we expect similar results for other types of tests.

In this paper, we identified the scale of the factors by fixing their variances to unity in each cluster and eliminated the rotational freedom of EFA by imposing the minimally required number of restrictions, which (in the case of two orthogonal factors) implies restricting one loading to be zero in each of the clusters. An alternative approach selects a representative variable (marker) for each factor (Little et al., 2006) and fixes the loading of the markers on the factor they measure to 1.0 and the loadings on the remaining (possibly correlated) factors to zero. Although the way in which the scale of the latent variable is identified does not impact model fit, the standard errors and the corresponding conclusions regarding the statistical significance of the estimated loading differences may vary based on the selection of the marker, or may differ when the scale of the latent variable is defined by fixing its variance (Brown, 2014). For the same reason the choice of the restricted zero loadings for eliminating rotational freedom should be an informed one (Millsap, 2001). In the simulation study, zero loadings and factor variances were known, but in empirical practice, a poor choice of identification restrictions affects the hypothesis testing. With respect to the scale of the factor, this problem is not specific to MSFA, but inherent to all mixture and multigroup factor analysis methods (e.g., Cheung & Rensvold, 1999; Millsap, 2001). With respect to the rotation, MSFA shares this problem with similar EFA-based methods. In case of more factors and/or more clusters, the choice of identification restrictions surely becomes very complex, because more

[5]This was established for heterogeneity across groups in MGFA.
[6]See footnote 5.

loadings need to be restricted and it may be unrealistic to assume that the set of restricted loadings and/or the scale of the factors are the same across clusters. Clearly, it is advantageous to make this choice obsolete in the future. On top of that, in empirical analyses, it is often desirable to rotate factor loadings toward structures that are easy to interpret for each group (simple structure, Thurstone, 1947) by means of rotational criteria (e.g., Varimax, Kaiser, 1958; oblimin, Jennrich, 1979). As a result, many loadings will approach zero, but the user does not need to specify which ones, and Wald tests apply for all the loadings because none of them are restricted. Additionally, for the sake of the Wald tests, it would be interesting to also rotate to maximal agreement of the factor loadings between groups to minimize the number of false positives. Note that a multigroup rotation procedure was recently proposed that solves rotational freedom as well as identification of factor scales such that hypothesis testing for loading differences is optimized (De Roover & Vermunt, 2019). This approach can be applied to MSFA as well, but its results depend on the choice of a rotational criterion, however, which was beyond the scope of the simulation study presented in this paper.

The last crucial issue is related to the multiple sources of factor loading differences. In this paper, we preferred to capture all differences in covariance structure by means of the clusters and the cluster-specific factor loadings, but for evaluating measurement invariance it may be desirable to disentangle differences in factor loadings from differences in factor (co) variances by including group-specific factor (co)variances in the model (De Roover et al., in press).

Appendix

The following Latent GOLD syntax has been used in simulation study 1 and 2 to simulate data matrices with $K = 2$ clusters and $Q = 2$ factors per cluster.

```
//LG5.1//
version = 5.1
infile 'C:\Users\Documents\. . ....\data_start.
txt' quote = single
model
options
  algorithm
    tolerance=1e-008 emtolerance=1e-008
    emiterations=2500 nriterations=0;
  startvalues
    seed=0 sets=1 tolerance=1e-005 iterations
    =1;
  bayes
    categorical=1 variances=1 latent=1 poisson
    =1;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature nodes=10;
  missing exclude all;
  output
    iteration detail classification
    parameters=effect standarderrors
    probmeans=posterior
    profile bivariate residuals;
  outfile 'sim.txt' simulation;
variables
  case weight freq;
  groupid V1;
  dependent
    V2 continuous, V3 continuous, V4 continu-
    ous, V5 continuous, V6 continuous,
    V7 continuous, V8 continuous, V9 continu-
    ous, V10 continuous, V11 continuous,
    V12 continuous, V13 continuous, V14
    continuous;
  independent V1 nominal;
  independent Cluster nominal;
  latent
    F1 continuous,
    F2 continuous;
equations
//factor variances
  (1) F1;
  (1) F2;
//regression models for items
  V2 - V14 <- F1 | Cluster + F2 | Cluster;
//unique variances
  V2 - V14 | Cluster;
//starting values
  'startingvalues.txt'
end model
```

The following Latent GOLD syntax was employed in the empirical application and in simulation study 1 to estimate an MSFA model with $K = 2$ clusters and $Q = 2$ factors per cluster for the empirical and simulated data sets, respectively. Rotational freedom was eliminated by restricting the secondary loading of *jealousy* (labeled here as V6) to zero in both clusters.

```
options
  algorithm
    tolerance=1e-008 emtolerance=1e-006
    emiterations=2500 nriterations=500;
  startvalues
    seed=0 sets=25 tolerance=1e-005
    iterations=100 PCA;
  bayes
    categorical=0 variances=0 latent=0 poisson
    =0;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature nodes=10;
  missing exclude all;
  output
    iterationdetail classification
    parameters=effect standarderrors
    probmeans=posterior profile
    bivariate residuals
    writeparameters='results_parameters.csv'
    write='results.csv'
    writeloadings='results_loadings.txt';
  outfile
    'classif.txt' classification;
variables
  groupid V1;
```

```
  dependent
    V2 continuous, V3 continuous, V4 continu-
    ous, V5 continuous, V6 continuous,
    V7 continuous, V8 continuous, V9 continu-
    ous, V10 continuous, V11 continuous,
    V12 continuous, V13 continuous, V14
    continuous;
  independent V1 nominal;
  latent
    F1 continuous,
    F2 continuous,
    Cluster nominal group 2 coding=first;
equations
//factor variances
  (1) F1;
  (1) F2;
//logistic regression model for clusters
  Cluster <- 1;
//regression models for items
  V2 - V5 V7 - V14 <- F1 | Cluster + F2 | Cluster;
  V6 <- (0) F1 | Cluster + F2 | Cluster;
//unique variances
  V2 - V14 | Cluster;
```

The following Latent GOLD syntax was employed in simulation study 2 to estimate an MSFA model with $K = 2$ clusters and $Q = 2$ factors per cluster. Rotational freedom was eliminated by restringing to zero the secondary loading of V14 in both clusters.

```
options
  algorithm
    tolerance=1e-008 emtolerance=1e-006
    emiterations=2500
nriterations=500;
  startvalues
    seed=0 sets=100 tolerance=1e-005
    iterations=100 PCA;
  bayes
    categorical=0 variances=0 latent=0 poisson
    =0;
  montecarlo
    seed=0 replicates=500 tolerance=1e-008;
  quadrature nodes=10;
  missing exclude all;
  output
    iterationdetail classification
    parameters=effect standarderrors
    probmeans=posterior profile
    bivariate residuals
    writeparameters='results_parameters.csv'
    write='results.csv'
    writeloadings='results_loadings.csv';
  outfile 'classif.txt' classification;
variables
  groupid V1;
  dependent
    V2 continuous, V3 continuous, V4 continu-
    ous, V5 continuous, V6 continuous,
    V7 continuous, V8 continuous, V9 continu-
    ous, V10 continuous, V11 continuous,
    V12 continuous, V13 continuous, V14
    continuous;
  independent V1 nominal;
  latent
    F1 continuous,
    F2 continuous,
    Cluster nominal group 2 coding=first;
equations
//factor variances
  (1) F1;
  (1) F2;
//logistic regression model for clusters:
  Cluster <- 1;
//regression models for items
  V2 - V13 <- F1 | Cluster + F2 | Cluster;
  V14 <- (0) F1 | Cluster + F2 | Cluster;
//unique variances
  V2 - V14 | Cluster;
```

Funding

The research leading to the results reported in this paper was partly funded by the Netherlands Organization for Scientific Research (NWO) [Veni grant 451-16-004].

ORCID

Elena Geminiani http://orcid.org/0000-0001-5992-9728
Eva Ceulemans http://orcid.org/0000-0002-7611-4683
Kim De Roover http://orcid.org/0000-0002-0299-0648

References

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

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A bayesian approach. *Psychometrika*, *67*(1), 49–77. https://doi.org/10.1007/BF02294709

Asparouhov, T. (2007). *Wald test of mean equality for potential latent class predictors in mixture modeling. technical appendix*. Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 92–108). The Guilford Press.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507–526. https://doi.org/10.1037/met0000077

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Bonferroni, C. E. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Libreria internazionale Seeber.

Breusch, T. S., & Pagan, A. R. (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, *47*(1), 239–253. https://doi.org/10.2307/2297111

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Buse, A. (1982). The likelihood ratio, Wald, and lagrange multiplier tests: An expository note. *The American Statistician*, *36*(3a), 153–157. https://doi.org/10.1080/00031305.1982.10482817

Buzick, H. M. (2010). Testing for heterogeneous factor loadings using mixtures of confirmatory factor analysis models. *Frontiers in Psychology*, *1*, 165. https://doi.org/10.3389/fpsyg.2010.00165

Chatfield, C. (1995). *Problem solving: A statistician's guide* (2nd ed.). Chapman & Hall/CRC Press.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. https://doi.org/10.1177/014920639902500101

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

De Raad, B., Barelds, D. P., Levert, E., Ostendorf, F., Mlačić, B., Blas, L. D., Hřebíčková, M., Szirmák, Z., Szarota, P., Perugini, M., Church, A. T., & Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology*, *98*(1), 160–173. https://doi.org/10.1037/a0017184

De Roover, K., Ceulemans, E., Timmerman, M. E., Vansteelandt, K., Stouten, J., & Onghena, P. (2012). Clusterwise simultaneous component analysis for analyzing structural differences in multivariate multiblock data. *Psychological Methods*, *17*(1), 100–119. https://doi.org/10.1037/a0025385

De Roover, K., Timmerman, M. E., & Ceulemans, E. (2017). How to detect which variables are causing differences in component structure among different groups. *Behavior Research Methods*, *49*(1), 216–229. https://doi.org/10.3758/s13428-015-0687-8

De Roover, K., Timmerman, M. E., De Leersnyder, J., Mesquita, B., & Ceulemans, E. (2014). What's hampering measurement invariance: Detecting non-invariant items using clusterwise simultaneous component analysis. *Frontiers in Psychology*, *5*, 604. https://doi.org/10.3389/fpsyg.2014.00604

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*(6), 905–923. https://doi.org/10.1080/10705511.2019.1590778

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

De Roover, K., Vermunt, J. K., Timmerman, M. E., & Ceulemans, E. (2017). Mixture simultaneous factor analysis for capturing differences in latent variables between higher level units of multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 506–523. https://doi.org/10.1080/10705511.2017.1278604

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: A Multidisciplinary Journal*, *16*(2), 295–314. https://doi.org/10.1080/10705510902751416

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, *65*(3), 457–483. https://doi.org/10.1093/biomet/65.3.457

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter-and intranational differences. *Journal of Personality and Social Psychology*, *81*(5), 869–885. https://doi.org/10.1037/0022-3514.81.5.869

Engle, R. F. (1984). Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handbook of Econometrics*, *2*, 775–826. https://doi.org/10.1016/S1573-4412(84)02005-5

Eysenck, H. J., & Eysenck, S. B. G. (1975). Manual of the Eysenck Personality Questionnaire (Junior and Adult) [Computer software manual]. Hodder & Stoughton.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research*, *27*(3), 387–415. https://doi.org/10.1207/s15327906mbr2703_5

Foulkes, A. S. (2009). *Applied statistical genetics with R: For population-based association studies*. Springer Science & Business Media.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, *59*(6), 1216–1229. https://doi.org/10.1037/0022-3514.59.6.1216

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

Gudicha, D. W., Tekle, F. B., & Vermunt, J. K. (2016). Power and sample size computation for Wald tests in latent class models. *Journal of Classification*, *33*(1), 30–51. https://doi.org/10.1007/s00357-016-9199-1

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The multigroup common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement*, *30*(3), 233–246. https://doi.org/10.1177/0146621605279760

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202–226. https://doi.org/10.1177/0013164404267287

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

Jennrich, R. I. (1979). Admissible values of γ in direct oblimin rotation. *Psychometrika*, *44*(2), 173–177. https://doi.org/10.1007/BF02293969

Jöreskog, K. G. (1967). A general approach to confirmatory maximum likelihood factor analysis. *ETS Research Bulletin Series*, *1967*(2), 183–202. https://doi.org/10.1002/j.2333-8504.1967.tb00991.x

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. https://doi.org/10.1007/BF02289233

Kiers, H. A. L. (1997, 12). Three-mode orthomax rotation. *Psychometrika*, *62*(4), 579–598. https://doi.org/10.1007/BF02294644

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. https://doi.org/10.1080/10705511.2017.1304822

Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., & Kim-Prieto, C. (2006). Universal intracultural and intercultural dimensions of the recalled frequency of emotional experience. *Journal of Cross-cultural Psychology*, *37*(5), 491–515. https://doi.org/10.1177/0022022106290474

Lawley, D. N., & Maxwell, A. E. (1962). Factor analysis as a statistical method. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *12*(3), 209–229. https://doi.org/10.2307/2986915

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with non-experimental data* (Vol. 53). John Wiley & Sons.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(1), 59–72. https://doi.org/10.1207/s15328007sem1301_3

Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*(2), 57–64. https://doi.org/10.1027/1614-2241.2.2.57

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

Ma, X., & Klinger, D. A. (2000). Hierarchical linear modelling of student and school effects on academic achievement. *Canadian Journal of Education*, *25*(1), 41–55. https://doi.org/10.2307/1585867

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139. https://doi.org/10.1207/S15327906MBR3801_5

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 611–635. https://doi.org/10.1080/10705510701575461

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

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. SAGE Publications.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 1–17. https://doi.org/10.1207/S15328007SEM0801_1

Molenaar, D., & Dolan, C. V. (2018). Nonnormality in latent trait modelling. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley Handbook of psychometric testing* (pp. 347–373). John Wiley & Sons, Ltd.

Mulaik, S. A. (2009). *Foundations of factor analysis*. Chapman and Hall/CRC.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes*, *17*, 1–48. https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B., & Asparouhov, T. (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, *34*(6), 1041–1058. https://doi.org/10.1002/sim.6388

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*(3), 241–286. https://doi.org/10.1006/ceps.2000.1040

Pennell, R. (1968). The influence of communality and N on the sampling distributions of factor loadings. *Psychometrika*, *33*(4), 423–439. https://doi.org/10.1007/BF02290161

R Core Team. (2020). *R: A Language and environment for statistical computing* [Computer software manual]. http://www.R-project.org/

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 44, pp. 50–57).

Rodriguez, C., & Church, A. T. (2003). The structure and personality correlates of affect in Mexico: Evidence of cross-cultural comparability using the Spanish language. *Journal of Cross-cultural Psychology*, *34*(2), 211–230. https://doi.org/10.1177/0022022102250247

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

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.

Sörbom, D. (1989). Model modification. *Psychometrika*, *54*(3), 371–384. https://doi.org/10.1007/BF02294623

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, *59*(2), 301–320. https://doi.org/10.1348/000711005X67599

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Tech. Rep.). Educational Testing Service Princeton NJ.

Vermunt, J. K., & Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, advanced, and syntax*. Belmont, MA: Statistical Innovations.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482. https://doi.org/10.1090/S0002-9947-1943-0012401-3

Wickrama, K. A. S., Lee, T. K., O'Neal, C. W., & Lorenz, F. O. (2016). *Higher-order growth curves and mixture modeling with mplus: A practical guide*. Routledge.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62. https://doi.org/10.1214/aoms/1177732360

Zhao, J., Yu, P., & Shi, L. (2013). *Model selection for mixtures of factor analyzers via hierarchical bic*. School of Statistics and Mathematics, Yunnan University of Finance and Economics.