STRUCTURAL EQUATION MODELING
2025, VOL. 32, NO. 1
ISSN: 1070-5511 | E-ISSN: 1532-8007

New Developments in Measurement Invariance Testing: An Overview and Comparison of EFA-Based Approaches

Philipp Sterner[a,b,c], Kim De Roover[d] and David Goretzko[a,c]

[a]LMU Munich; [b]German Center for Mental Health (DZPG); [c]Utrecht University; [d]KU Leuven

ABSTRACT

When comparing relations and means of latent variables, it is important to establish measurement invariance (MI). Most methods to assess MI are based on confirmatory factor analysis (CFA). Recently, new methods have been developed based on exploratory factor analysis (EFA); most notably, as extensions of multi-group EFA, researchers introduced mixture multi-group EFA, multi-group exploratory factor alignment, EFA trees, and multi-group factor rotation to resolve rotational indeterminacy in EFA. The main advantage of EFA-based (compared to CFA-based) assessment of MI is that no potentially too restrictive measurement model has to be specified. This allows for a more thorough investigation because violations of MI due to cross-loadings can be considered, too. For each method, we address the model specification and recommendations for application, detailing their strengths and weaknesses. We demonstrate each method in combination with multi-group factor rotation in an empirical example. Differences to and possible combinations with CFA-based methods are discussed.

KEYWORDS

Alignment; cross-loadings; exploratory factor analysis; factor rotation; measurement invariance

In psychological science, we are almost always interested in investigating some kind of latent variable (e.g., personality traits like extraversion). The object of research is often the comparison of mean values of latent variables (or measurements thereof) between different groups; for example, prosociality or moral judgements across countries (Bago et al., 2022; House et al., 2020). This includes both comparisons between different groups (e.g., in cross-cultural research; Milfont & Fischer, 2010) or comparisons across subsequent measurements within the same group (e.g., pre- and post-treatment). Latent variables are measured by so called indicators or observed variables (often questionnaire items) in order to obtain scores of the latent variable (Lord & Novick, 1968; Van Bork et al., 2022). The relationship between observed and latent variables is captured in the *measurement model*. To enable meaningful comparisons between groups, it is crucial to test whether the measurement models are invariant across groups. *Measurement invariance* (MI) means that the latent variables are measured identically across groups; that is, people with the same true score on the latent variable should also receive the same score on the observed variables (Meredith, 1993; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). In more technical terms, the parameters of the measurement model have to be identical across groups.

Multi-group Confirmatory Factor Analysis (MG-CFA) was originally introduced to test whether a measurement model is invariant across a defined set of groups. However, MG-CFA reaches its limits when many groups have to be compared (e.g., a covariate *nation* with 48 groups; Kuppens et al., 2006). The chance of false-positive findings of non-invariance increases with the number of groups due to multiple testing (Rutkowski & Svetina, 2014). Additionally, this amount of hypothesis tests can make it difficult to tell invariant from non-invariant parameters (Byrne & Vijver, 2010; De Roover et al., 2022). To improve investigations of MI for cases with many groups, more advanced methods have been developed. Raykov et al. (2013) developed a multiple testing procedure to investigate MI that uses the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). This controls the false-discovery rate rather than the family-wise error rate, resulting in a higher power compared to simple multiple testing of MG-CFAs with Bonferroni correction. Kim et al. (2017) provide a comprehensive overview of methods to investigate MI with many groups, for example, Multilevel Factor Mixture Modeling and Alignment Optimization (Asparouhov & Muthén, 2014). The majority of these methods developed so far and all methods detailed in Kim et al. (2017) are based on CFA. While the use of CFA allows to incorporate theoretical considerations when investigating MI, it can also be too restrictive in terms of model specification. If the model is slightly misspecified, a CFA-based approach might not accurately recover the true structure of a model (Nájera et al., 2023).

In recent years, new methods have been developed that are based on *exploratory factor analysis* (EFA). As extensions of *multi-group EFA* (MG-EFA; Dolan et al., 2009),

researchers developed *mixture multi-group EFA* (MMG-FA; De Roover et al., 2022), *multi-group exploratory factor alignment*[1] (AESEM; Asparouhov & Muthén, 2023), and *EFA trees* (Sterner & Goretzko, 2023). Investigating MI on the basis of EFA avoids the problem of having to assume a (potentially too) restrictive model across all groups. The goal of this paper is to give an overview of these recent developments and to demonstrate the application of EFA-based MI methods. Additionally, because all of these methods inherit the challenge of rotational indeterminacy of the EFA model, we illustrate how these methods can be combined with *multi-group factor rotation* (MGFR; De Roover & Vermunt, 2019). MGFR resolves the rotational indeterminacy per group and locates non-invariant factor loadings by means of hypothesis testing. We demonstrate all of this on an empirical data example from moral psychology (Bago et al., 2022). Because the methods differ in their assumptions and outcomes, a direct comparison does not make too much sense. Instead, we provide a guide on when to use which method in which way. By this, we hope to help researchers to navigate through the extensive literature on EFA-based methods to investigate MI and increase the prevalence of MI testing in social scientific research (Leitgöb et al., 2023; Maassen et al., 2023). To facilitate the application of the presented methods, we provide openly available *R*, *Mplus*, and *Latent Gold* code.

The remainder of the paper is structured as follows: Section 1 outlines the differences between CFA- and EFA-based tests of MI. Sections 2–5 present the four EFA-based methods (MG-EFA, MMG-FA, AESEM, EFA trees) in detail. For each method, we address the model specification and recommendations on when to use the method, detailing their strengths and weaknesses. Section 6 presents MGFR and an overview table summarizing all methods. Sections 7 and 8 demonstrate the application of the methods in combination with MGFR. Section 9 discusses differences to and possible combinations with CFA-based methods.

1. CFA vs. EFA in MI Testing

The main difference between CFA and EFA pertains to the loadings in the model. A loading quantifies the strength of the relation between a latent factor and an item. In CFA, some loadings are constrained to zero whereas in EFA all paths between latent and observed variables are estimated freely (Goretzko et al., 2021; Mulaik, 2010). Thus, if assumptions about which items measure which latent factor are available, CFA allows to incorporate these assumptions in the model. If there are no assumptions and the goal is to uncover the relation between items and latent factors, EFA should be preferred. This is especially the case during the development of new measures.

[1]We abbreviate the method by AESEM following Asparouhov and Muthén (2023). They extended the alignment method (Asparouhov & Muthén, 2014) to the general *exploratory structural equation model* (ESEM; Asparouhov & Muthén, 2009), which leads to the abbreviation AESEM (aligned ESEM). However, we will only look at the measurement model part of ESEMs, which are EFAs.

As already mentioned, until recently, CFA was the basis for most MI testing methods (Marsh et al., 2014). As a consequence, MI in this context not only concerns the equivalence of parameters in the measurement model but also the equivalence of its architecture; that is, the number of latent factors and the imposed zero-loadings must hold across groups (De Roover et al., 2022). Needless to say, the strict specification of the measurement model with zero-loadings is often not tenable (Nájera et al., 2023), especially when these restrictions have to be assumed across all groups. If the model is then modified in a data-driven way, its generalizability is diminished because this strategy capitalizes on chance (MacCallum et al., 1992). Additionally, misspecifications in the measurement model can introduce bias in the estimation of the remaining parameters, especially when maximum likelihood estimation is used (Bollen et al., 2007). Since in EFA no zero-loadings are imposed, none of these problems caused by model misspecifications are an issue in EFA-based MI testing. EFA as a basis even makes tests for MI wider-ranging because it allows to assess the invariance of cross-loadings as well as differences in the position of main loadings (De Roover & Vermunt, 2019). These advantages are inherent in all methods that we will present.

2. Multi-Group EFA

2.1. Model Specification

Both MG-CFA and MG-EFA are instances of the more general multi-group factor analysis model (Jöreskog, 1971; Sörbom, 1974). In MG-EFA, no loading paths between the observed variables and the latent factors are constrained to zero. Hence, EFA can be used to freely uncover the relations between observed and latent variables (Goretzko et al., 2021). Let $\boldsymbol{x}_{i_g}$ be the p-dimensional vector of observed variables for subject i_g in group g (with $i_g = 1, \ldots, N_g$ and $g = 1, \ldots, G$). This vector can be described as a linear function of the m latent factors (Mulaik, 2010):

$$\boldsymbol{x}_{i_g} = \boldsymbol{\tau}_g + \boldsymbol{\Lambda}_g \boldsymbol{\xi}_{i_g} + \boldsymbol{\epsilon}_{i_g} \tag{1}$$

where $\boldsymbol{\tau}_g$ is a p-dimensional vector of group-specific intercepts, $\boldsymbol{\Lambda}_g$ is a $p \times m$ matrix of group-specific factor loadings, $\boldsymbol{\xi}_{i_g}$ is a m-dimensional vector of latent factor scores, and $\boldsymbol{\epsilon}_{i_g}$ is a p-dimensional vector of error terms. For maximum-likelihood estimation, the latent factor scores are assumed to be multivariate-normally distributed; specifically, $\boldsymbol{\xi}_{i_g} \sim MVN(\boldsymbol{\alpha}_g, \boldsymbol{\Phi}_g)$, where $\boldsymbol{\alpha}_g$ denotes the factor means of group g and $\boldsymbol{\Phi}_g$ the factor (co-)variances. In MG-EFA, the factors are rotationally indeterminate per group, which means there are infinitely many sets of factor solutions which have the same fit to the data but lead to different interpretations of the solution. This has to be resolved per group by a rotation criterion (De Roover & Vermunt, 2019), which often improves interpretability by pursuing simple structure – where each observed variable has near-zero loadings for all factors but one. As already mentioned, we will employ MGFR to address this issue (details will follow in a later section), which not only strives for simple structure but also maximizes the similarity of the rotated loadings across groups. The error terms are also assumed

to be multivariate-normal and independent of the factor scores; specifically, $\boldsymbol{\epsilon}_{i_g} \sim MVN(0, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\Psi}_g$ is a $p \times p$ diagonal matrix which contains the unique variances of the observed variables in group g. Combining all of the above, we arrive at the group-specific model-implied covariance matrix $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g^\top + \boldsymbol{\Psi}_g$.

2.2. Testing Procedure

In a factor-analytic context, MI is tested by fitting and comparing increasingly constrained models (De Roover et al., 2022; Vandenberg & Lance, 2000). Many comprehensive guides for MI testing in this context exist, so we will keep this section rather short (see e.g., Putnick & Bornstein, 2016; Van de Schoot et al., 2012). The first step is to test whether configural MI holds. Configural MI means that the construct architecture, that is, the number of latent factors and the location of zero-loadings are equivalent across groups. This is tested by estimating the baseline model in Equation (1) per group. Because there are no loadings constrained to zero in MG-EFA, the only model misspecification that could cause the overall model fit to be bad is a different number of latent factors. For example, the baseline models for all groups are estimated with three latent factors but in one group there are actually four latent factors. To partially identify the model, the factor means $\boldsymbol{\alpha}_g$ are set to 0 and the factor covariance matrix $\boldsymbol{\Phi}_g$ is set to an $m \times m$ identity matrix, that is, with factor variances of 1 and factor covariances of 0 (Van de Schoot et al., 2012). In the next step, the invariance of factor loadings, called weak or metric MI, is tested. For this, the fit of the baseline model is compared to the fit of a model in which loadings are constrained to be equal across groups (i.e., $\boldsymbol{\Lambda}_1 = \ldots = \boldsymbol{\Lambda}_G$). If metric MI is supported, latent covariances or relations (e.g., how extraversion relates to other latent variables) can be compared between groups (De Roover et al., 2022). Strong or scalar MI is assessed by comparing the fit of the metric model with the fit of a model with constrained intercepts (i.e., $\boldsymbol{\tau}_1 = \ldots = \boldsymbol{\tau}_G$). If scalar MI is supported, comparisons of latent factor means are warranted (e.g., the means of extraversion). In the last step, strict or residual MI is tested by constraining the unique variances of the observed variables, that is, the diagonal of $\boldsymbol{\Psi}_g$, to be equal across groups. If residual MI holds and factor variances are equal as well, this means that the item reliabilities are equal across groups (e.g., extraversion is measured with the same precision in different groups) (Vandenberg & Lance, 2000). However, this level of MI can be difficult to achieve and is not a prerequisite for the comparison of latent factor means (Chen, 2007; Vandenberg, 2002).

A decrease in fit when estimating a more restricted model is an indication that the tested level of MI is not supported (Chen, 2007; Cheung & Rensvold, 2002); for example, if the comparative fit index (CFI) decreases by more than 0.01 and/or the root mean squared error of approximation (RMSEA) increases by more than 0.01. Rutkowski and Svetina (2014) propose more liberal cut-offs for when the number of groups exceeds 10, especially for testing metric MI: a decrease of the CFI by more than 0.02 and an increase of the RMSEA by more than 0.03. Appropriate cutoffs for model fit evaluation depend on both model complexity and sample size (Cao & Liang, 2022b; Goretzko et al., 2023), so researchers should not carelessly adopt proposed values. Cao and Liang (2022a) provide more detailed recommendations on the choice of common fit measures to detect violations of MI in models with cross-loadings. A stricter comparison of the models by a χ^2-difference test is also possible as the respective models are always nested. However, as the test is highly sensitive to sample size, the use of fit indices is widely considered more suitable (De Roover et al., 2022).

2.3. When to Use MG-EFA

When applying MG-EFA, no statistical knowledge beyond that of single-group EFA is needed. Instead of investigating one loading matrix, researchers get to work with up to G loading matrices (with G being the number of investigated groups). One thing that is more challenging in MG-EFA is the choice of rotation and its interpretation. Choosing the right rotation is never easy because, depending on the rotation, different interpretations of the factor solutions emerge. When dealing with more than one loading matrix, the conclusions about invariance or non-invariance might change when using different rotations (De Roover & Vermunt, 2019). The issue of rotation in the multi-group case will be discussed thoroughly in the section on MGFR. It should also be kept in mind that whereas in the single-group case the data are usually standardized, in multi-group settings it is common to use unstandardized data (i.e., to model covariance instead of correlation matrices).

MG-EFA comes with a lack of flexibility and strong assumptions that have to be made. MG-EFA can only test MI on covariates that are measured and for which hypotheses about non-invariance exist. If a covariate associated with non-invariance is not measured or if there are no hypotheses about non-invariant group constellations, MG-EFA reaches its limits. For example, researchers have to choose the covariate *gender* and form hypotheses about non-invariant groups to test MI on this covariate. However, more often than not the choice of which covariate to test for non-invariance is not straightforward (Sterner et al., 2024). Testing all available covariates with all potential group constellations leads to the already mentioned multiple testing problem. This is emphasized in cases where a covariate encompasses many groups and nearly impossible when a covariate is continuous (e.g., *age*; Putnick & Bornstein, 2016). To summarize, if you want to test MI for a measured categorical covariate with a small number of groups, and if specifying a CFA model might be too strict, MG-EFA is a good option. If not, you might want to resort to one of the methods presented in the following. We will explain how these methods can find unmeasured clusters of groups for which MI holds, investigate MI along a continuous covariate or identify covariates associated with MI without any hypotheses about them.

3. Mixture Multi-Group EFA

3.1. Model Specification

MMG-EFA extends MG-EFA by building on the assumption that, although parameters differ across groups, some groups have equal measurement parameters. Thus, there may be *clusters of groups* based on these parameters for which MI is supported. Therefore, MMG-EFA performs clustering based on finite mixtures (McLachlan et al., 2019) to identify groups that have equal parameters in the measurement model (Leitgöb et al., 2023), for example, equal loadings (De Roover et al., 2022) and/or equal intercepts (De Roover, 2021). Groups within the same cluster are then modeled with cluster-specific loadings and/or intercepts. Parameters of the measurement model that pertain to a higher level of invariance (e.g., unique variances) are still estimated group-specifically. The parameters of the structural model (i.e., factor means and factor (co)variances) are also free to vary among groups in the same cluster. The assumption of underlying clusters implies that the data-generating model of the observed variables $\boldsymbol{x}_{i_g}$ is a mixture of multivariate-normal distributions with K components (which we call clusters). All observations of a group are assumed to stem from the same normal distribution, that is, there are no parameter differences below the group-level (e.g., differences on the observation-level within a group). Because EFA-based methods are especially useful for evaluating main- and cross-loading differences between groups, we focus on the model with cluster-specific loadings (De Roover et al., 2022). We refer readers interested in the model with cluster-specific intercepts to De Roover (2021). The MMG-EFA model with cluster-specific loadings for group g is

$$f(\boldsymbol{X_g};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{gk}(\boldsymbol{X_g};\boldsymbol{\theta}_{gk}) = \sum_{k=1}^{K} \pi_k \prod_{i_g=1}^{N_g} MVN(\boldsymbol{x}_{i_g};\boldsymbol{\mu_g},\boldsymbol{\Sigma}_{gk}) \tag{2}$$

where the model-implied covariance matrix for group g conditional on the cluster membership $z_{gk} = 1$ is given by $\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Lambda_k}\boldsymbol{\Phi}_{gk}\boldsymbol{\Lambda_k}^{\top} + \boldsymbol{\Psi_g}$. Note that $\boldsymbol{\mu_g}$ are the group-specific item means, which are equal to the intercepts in case of factor means of 0. The density of the distribution of the whole population is denoted by f. The prior classification probability of a group to belong to each of the K clusters is indicated by π_k (thus, $\sum_{k=1}^{K} \pi_k = 1$) and f_{gk} is the k^{th} cluster-specific density for group g. $\boldsymbol{\theta}_{gk}$ denotes the parameter set of these distributions, containing both the mean vectors and covariance matrices. After model estimation, posterior classification probabilities $\hat{z}_{gk}$ are obtained that indicate the estimated probability that group g belongs to cluster k. Notice how the loading matrices $\boldsymbol{\Lambda_k}$ are now cluster-specific (with rotational freedom *per cluster*), whereas the intercepts $\boldsymbol{\tau_g}$ and the unique variances $\boldsymbol{\Psi_g}$ remain group-specific (for an explanation why the factor (co-)variances $\boldsymbol{\Phi}_{gk}$ are group- and cluster-specific, see De Roover et al., 2022). This renders the covariance matrix $\boldsymbol{\Sigma}_{gk}$ to be group- *and* cluster-specific but only the cluster-specific loadings influence the clustering.

It is important to note that the invariance of parameters within each cluster only holds under the assumption that the correct number of clusters K was extracted. If too few clusters are selected, MI may not hold within each cluster. If too many clusters are selected, MI may hold across some of the clusters. This model selection problem is addressed by combining both the *Bayesian Information Criterion* (BIC; Schwarz, 1978) and the *Convex Hull procedure* (Ceulemans & Kiers, 2006; CHull; Ceulemans & Van Mechelen, 2005). The BIC tries to strike a balance between model fit and complexity by adding a penalty for additional free parameters and larger sample sizes. De Roover et al. (2022) and De Roover (2021) recommend to use the number of groups G for the sample size when computing the BIC (instead of the actual sample size) because the clustering operates at the group level. CHull can be seen as a generalization of the scree test (Cattell, 1966), again trying to balance model fit and complexity. This is similar to the approach suggested by Lorenzo-Seva et al. (2011) to determine the number of factors to be extracted in EFA. We refer readers to De Roover et al. (2022) and De Roover (2021) for more technical details on these two model selection strategies. It is best to run multiple MMG-EFAs with different numbers of clusters and to choose the solution with the lowest BIC and the highest scree-ratio resulting from CHull. In general, when in doubt about how many clusters to extract, it is recommended to investigate the two or three best solutions. Only if the additional clusters show substantive parameter differences, the solution with more clusters should be preferred over the parsimonious solution with less clusters (De Roover, 2021). Additionally, applying MG-EFA per cluster to test whether MI holds within each cluster can be a way to check if the selected number of clusters is plausible.

3.2. When to Use MMG-EFA

As mentioned, we focus on the MMG-EFA model with cluster-specific loadings (De Roover et al., 2022) but the following points also apply to the model with cluster-specific intercepts (De Roover, 2021). MMG-EFA proves especially useful when you want to efficiently investigate measurement (non-)invariance across many groups. By introducing the assumption that there are clusters of invariant groups, the number of parameters that have to be compared in a pairwise manner are reduced. This can even be beneficial in case of a medium number of groups. For example, in the case of six groups, 15 pairwise comparisons would be needed to test all possible pairs of groups for MI. By assigning these six groups to three clusters (two groups each), the number of pairwise comparisons is reduced to three. Needless to say, the higher the number of comparisons, the higher the risk of falsely detecting non-invariance (De Roover, 2021; Rutkowski & Svetina, 2014).

Similarly, finding clusters of groups according to their measurement parameters helps in pinpointing which items are problematic with regard to MI (De Roover, 2021). By comparing the cluster-specific loadings or intercepts, items that are the source of non-invariance can be identified, again with less

pairwise comparisons. Another advantage is that the clustering might help to remedy small group sizes. When group sizes are too small to allow for a precise estimation of group-specific parameters, estimating parameters (e.g., loadings) cluster-specifically helps to achieve more reliable estimates.

For now, MMG-EFA can only be applied to continuous data, for which the assumption of normality is plausible. At least, data should be ordinal with five or more answer categories and no severe non-normality (De Roover et al., 2022). Researchers should thus make this assumption deliberately and should ensure that the data are approximately normal. Consequently, checking whether the data are approximately normal (ideally per group, since normality is assumed per cluster), having at least five answer categories for the questionnaire items, and having a large sample (to mitigate the effects of non-normality) are recommended when applying MMG-EFA (De Roover et al., 2022; Dolan, 1994).

4. EFA Trees

4.1. Model Specification

Usually, MI is tested with regard to the covariate of interest for comparison (e.g., gender). However, MI could also be violated in a more nuanced way by another covariate which is not considered. EFA trees can uncover covariates that are associated with violations of MI in a data-driven manner, that is, without any prior assumptions about which covariates to investigate (Sterner & Goretzko, 2023). To do so, they make use of model-based recursive partitioning (Hothorn et al., 2006; Zeileis et al., 2008). This algorithm tests whether parameters of the model are stable across groups that are defined by some covariate. If the parameters are unstable, it splits the data on the covariate which best explains this instability. More specifically, they loop through a three-stage process (Zeileis et al., 2008):

1. A model (in our case, an EFA) is fit to the entire sample by estimating the model parameters via maximum likelihood estimation. Let $\Pi(\boldsymbol{Y}, \boldsymbol{\theta})$ be the objective function, $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ the vector of model parameters (i.e., factor loadings, factor correlations, and unique variances) and $\boldsymbol{Y}$ the observations, with elements Y_i, $i = 1, ..., N$. The parameter estimates $\hat{\boldsymbol{\theta}}$ can be obtained by solving the first order condition

$$\sum_{i=1}^{N} \pi(Y_i, \hat{\boldsymbol{\theta}}) = 0 \tag{3}$$

whereby

$$\pi(\boldsymbol{Y}, \boldsymbol{\theta}) = \frac{\partial \Pi(\boldsymbol{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{4}$$

is the score function of $\Pi(\boldsymbol{Y}, \boldsymbol{\theta})$.
2. A test for parameter stability is performed with regard to every covariate by means of null hypothesis tests (*structural change test*). For this, the algorithm assesses whether the corresponding scores evaluated at the parameter estimates, $\hat{\pi}_i = \pi(Y_i, \hat{\boldsymbol{\theta}})$, fluctuate randomly around their mean 0. In each node, the model needs to be estimated only once to assess parameter stability (i.e., MI) with regard to different covariates. After every covariate has been evaluated, the one associated with the lowest (Bonferroni-corrected) p-value below a significance level α is selected for splitting the model. Note that by Bonferroni-correcting the p-values, the prespecified significance level α is ensured for the whole tree and the issue of multiple testing is accounted for (see Zeileis et al., 2008 for details on the distribution of the test statistics and how corresponding p-values are computed).
3. Once a covariate for splitting is found, the optimal split point on this covariate has to be computed. Note how the identification of a covariate to split on and the search for the split point on this covariate are two separate steps. This ensures that the bias of other tree algorithms (like CART or C4.5) toward selecting covariates with many potential split points is remedied. When splitting the model into B segments, two potential segmentations can be compared by evaluating the segmented estimation functions $\sum_{b=1}^{B} \sum_{i \in I_b} \Pi(Y_i, \boldsymbol{\theta}_b)$. For continuous covariates, an exhaustive search over all potential segmentations is performed. For a split into $B = 2$ segments, this can be performed in $O(N)$ operations, where N is the sample size. For categorical covariates, all potential constellations are evaluated. For a split into $B = 2$ segments, this can be performed in $O(2^{C-1})$ operations, with C being the number of categories. To keep the computational demand low and the examples illustrative, we only consider splits into two segments. However, this still allows us to identify covariates that define more than two non-invariant groups. For this, an EFA tree would simply split twice (or multiple times) on this covariate.

These three steps are repeated until a) no parameter instability in a leaf node is statistically significant, b) a prespecified depth of the tree is reached, or c) the sample size in a leaf node falls below a prespecified minimal value. For more mathematical details on the structural change tests, see Hothorn et al. (2006), Zeileis and Hornik (2007) and Zeileis et al. (2008). For more details regarding EFA trees specifically, see Sterner and Goretzko (2023).

4.2. When to Use EFA Trees

The main advantage of EFA trees is that no hypotheses about covariates potentially associated with (non-)invariance are needed. EFA trees automatically test all covariates for non-invariance, as opposed to (M)MG-EFA where grouping covariates have to be specified. In this, they can simultaneously handle categorical and continuous covariates. If one expects non-invariant groups associated with interactions between covariates, these interactions can be detected in two ways (Zeileis et al., 2008): Either the interaction term is added as a potential split covariate into the algorithm; or, to

preserve the exploratory spirit of EFA trees, one could allow "deeper" trees, that is, trees that split the data more than once. All splits in a tree are conditional on all prior splits. Suppose an EFA tree splits the data twice on two different covariates *age* and *gender*: Each leaf node (the final node in a tree) can be seen as a group defined by an interaction between these two covariates that lead to this leaf node, for example, women that are older than 30 years.

One issue that has to be kept in mind is that EFA trees are rather uninformative as to why they split the data. That is, there is no information available about which parameters of the measurement model differ across groups, causing the tree to split the data (Sterner & Goretzko, 2023). Consequently, researchers have to thoroughly investigate the models in the leaf nodes. This requires both domain expertise and experience in interpreting EFA results (e.g., different rotations of a loading matrix). As already mentioned, different rotations of the resulting solutions might lead to different conclusions about MI (De Roover & Vermunt, 2019). One remedy we will present below is the use of MGFR on the models in the nodes. Alternatively, Sterner and Goretzko (2023) describe how to apply elastic net regularization on the EFA models in the leaf nodes. Note that, given a specific type of regularization and set of hyperparameters, regularization yields a unique solution. However, changing these settings can again alter the conclusions about MI.

Even though EFA trees can assess MI on multiple covariates at the same time, it can only detect MI if the covariate causing it is measured (Sterner & Goretzko, 2023). If this covariate is not measured but a covariate correlated with the relevant one is available, non-invariance may still be detected (Strobl et al., 2015). As a consequence, if an EFA tree splits the data on a covariate, we would be cautious to interpret this covariate as the *cause* of the non-invariance. Every covariate identified for splitting could also be an observed indicator of a latent cause. Again, this underpins the importance of domain expertise when interpreting EFA trees.

To summarize, EFA trees require no hypotheses about the grouping variable(s) that is (are) relevant to capturing invariance and non-invariance in the data. However, domain expertise to interpret their results are indispensable. We recommend EFA trees for two scenarios specifically: First, in the earliest stages of questionnaire development, EFA trees allow for a thorough screening of various covariates and therefore numerous groups with varying measurement models. Even though MI is usually considered prior to latent mean comparisons, taking it into account when constructing a measure can help to prevent later issues with data analysis. The exploratory nature of EFA trees can assist researchers to consider every possible group constellation in this phase. Second, they can be applied prior to group comparisons with many available covariates, especially when many covariates are continuous. EFA trees can help to identify interactions that should be accounted for, in order to not render group comparisons meaningless.

5. Multi-Group Exploratory Factor Alignment

5.1. Model Specification

Alignment aims at enabling a comparison of latent means across groups when full MI is not supported; that is, when there are some small differences in parameters across groups (Asparouhov & Muthén, 2014, 2023). This is done by first estimating a configural model, that is, a model where all parameters are estimated group-specifically (this corresponds to the model in Equation (1)). The factor means and variances of these models are set to 0 and 1, respectively, for each group. In a second step, the alignment step, the factor means and variances of the groups are chosen so that the amount of non-invariance across groups is minimized. This corresponds to minimizing the differences between loadings and intercepts across groups. It is important to note that the factor means and variances are unidentifiable. As a consequence, the alignment does not change the model fit when searching for optimal values of the factor means and variances. To resolve this unidentifiability and to arrive at the optimal (i.e., "most invariant") values, an alignment function F is minimized with respect to the factor means and variances (Asparouhov & Muthén, 2014, 2023):

$$F = \sum_{m}\sum_{p}\sum_{g_k<g_l} w_{g_k,g_l} f(\lambda_{mpg_k} - \lambda_{mpg_l}) + \sum_{p}\sum_{g_k<g_l} w_{g_k,g_l} f(\tau_{pg_k} - \tau_{pg_l}) \quad (5)$$

where g_k and g_l represent groups k and l (with $k = l$) for every possible pair, and λ_{mpg_k} and λ_{mpg_l} (τ_{pg_k} and τ_{pg_l}) indicate the factor loadings (intercepts) of groups k and l, respectively. Because in AESEM cross-loadings are considered, all factors are aligned at the same time, in contrast to the original alignment method where all factors are aligned separately (Asparouhov & Muthén, 2023). w_{g_k,g_l} is a weight that depends on the group sizes, $w_{g_k,g_l} = \sqrt{N_{g_k} N_{g_l}}$, and thus expresses the certainty with which parameters for a group are estimated. The component loss function f is used to scale the observed parameter differences among the groups. It is chosen to be $\sqrt{\sqrt{x^2 + \epsilon}}$, where ϵ is a small number, e.g., 0.001. This function is approximately the same as $\sqrt{|x|}$ with ϵ being added to ensure continuous differentiability (Asparouhov & Muthén, 2014; Robitzsch, 2023). Equation (5) is minimized when the majority of loadings and intercepts are invariant, and only a small number of parameters are (largely) non-invariant (i.e., the number of non-invariant parameters is minimized). Medium-sized non-invariant parameters are avoided by this specific loss function (Kim et al., 2017).

Alignment cannot be considered a test of a specific level of MI (e.g., metric or scalar MI). However, the *Mplus* output provides invariance hypothesis tests for all parameters across groups (Flake & McCoach, 2018; Luong & Flake, 2023). That is, for every parameter estimate (e.g., for every loading), it is tested whether it is equivalent across groups. Additionally, an effect size estimate R^2 is provided for each parameter (Asparouhov & Muthén, 2014). This coefficient indicates the degree to which a parameter is invariant across groups, ranging from 0

(completely non-invariant) to 1 (completely invariant). The combination of these hypothesis tests and effect size estimates is an indication for the degree of (non-)invariance of a parameter across groups (Flake & McCoach, 2018).

The alignment approach has been extended to the EFA model in Asparouhov and Muthén (2023). The only difference to the procedure just described is that the (unrotated) configural model is rotated first, before being aligned. As usual, the rotation is done by minimizing a rotation criterion (e.g., geomin). Quite naturally, these separate steps of rotation and alignment can also be combined by adding the rotation function to the alignment loss function in Equation (5). This joint function is then minimized with respect to the factor means, factor variances, and the rotation criterion (i.e., usually a criterion aiming at simple structure solutions). In order to preserve the order of first rotating and then aligning the model, Asparouhov and Muthén (2023) assign an infinitely large weight to the rotation part of the joint function. As a consequence, the method first estimates a rotated configural model which is then aligned, conditional on the rotated solution.

5.2. When to Use AESEM

As already mentioned, AESEM—or alignment, in general—is not a test of MI but enables a comparison of latent means without having to make the assumption of exact MI. Especially in cases with many groups, there are many possibilities of MI being violated, so assuming exact MI is often unrealistic (Davidov et al., 2014). One assumption that has to be made for AESEM, however, is that most measurement parameters are invariant and only few parameters are non-invariant. A rough rule-of-thumb in the literature is that 25% of the parameters can be non-invariant (Asparouhov & Muthén, 2014; Flake & McCoach, 2018; Luong & Flake, 2023). If one is willing to make this assumption, AESEM produces a model with a clear interpretation about (non-)invariance. Researchers are provided with approximate latent means which can be used to compare groups even if exact MI is not supported (Asparouhov & Muthén, 2014, 2023). This makes AESEM a powerful follow-up method for the other methods presented here and elsewhere because it provides a way of handling non-invariant measurement models. One advantage of alignment in general is that it is well-researched under various conditions and that applications of the method on real data exist. For example, we refer readers to Munck et al. (2018) and Lomazzi (2018) for exemplary applications, to Luong and Flake (2023) for an in-depth tutorial, and to Flake and McCoach (2018) for a simulation study on its performance with polytomous items. Rudnev (2019) details a tutorial on alignment with *Mplus* syntax.

A disadvantage of alignment in combination with EFA (i.e., AESEM) is that the implemented rotation does not pursue agreement of loading matrices between groups. Instead, when rotating the configural models before the alignment step, AESEM solely applies a common rotation criterion like simple structure rotation (e.g., geomin, pursuing one non-zero loading per item) in every group (Asparouhov & Muthén, 2023). While this yields interpretable loading matrices *per group*, it is suboptimal for the evaluation of loading differences *between groups* (De Roover & Vermunt, 2019). As we will describe in the section on MGFR, a combined criterion that optimizes simple structure per group and agreement between groups would be a more suitable choice. Further, it is unfortunate that such a powerful method is only properly implemented in the commercial software *Mplus*. Many more researchers could benefit from this tool if open-source implementations were available. The only open-source implementation of the alignment function is provided in the the R package *sirt* (Robitzsch, 2022). However, because alignment is not the focus of the sirt package, its functionalities are limited compared to *Mplus* (e.g., it only supports alignment per factor, that is, for CFA models without cross-loadings).

6. Multi-Group Factor Rotation

6.1. Model Specification

As we have mentioned several times throughout this paper, EFA models are only determined up to admissible rotations. That is, the estimated solutions can be rotated in infinitely many ways without altering the goodness of fit of the model. To resolve this rotational indeterminacy, a rotation criterion has to be specified (see Browne, 2001 for an overview). What changes with different rotations, however, is the interpretation of the solutions. Depending on the rotation, the loading patterns (i.e., the size and probably also the allocation of primary and cross-loadings to latent factors) may completely change for a group. This, in turn, affects the conclusions regarding (non-)invariance across groups. Consequently, the choice of the rotation criterion is critical in a multi-group context (De Roover & Vermunt, 2019). If we are only interested in whether all loadings are invariant, the rotation of the solution is irrelevant. This is because a fully invariant factor model will show invariant loading patterns among all groups for every admissible rotation. We would then just compare the fit of the configural model and the metric MI model. If the loadings are invariant, we can impose equal loadings across groups and apply simple structure rotation or target rotation to this single set of loadings. If, however, loadings are non-invariant across group, we need to stick with the group-specific loadings and our goal would be to identify which loadings are non-invariant. This is needed to consider partial MI or item selection, or to reason about potential sources of non-invariance (De Roover & Vermunt, 2019). Solely applying simple structure rotation per group would not be optimal because it does not pursue agreement of the rotated factor loadings between groups. De Roover and Vermunt (2019) introduced MGFR to solve this rotation issue and provide a way of identifying loading differences between groups by means of hypothesis testing.[2] By applying MGFR, the solutions are rotated both to simple structure per group and to agreement between

[2]To enable hypothesis testing of rotated factor loadings, Jennrich (1973) showed how to derive the standard errors.

groups. For this, MGFR minimizes a rotation criterion and an agreement criterion (i.e., minimizes disagreement between groups) in a combined multi-group criterion:

$$R^{MG}(\mathbf{\Lambda_1}, \ldots, \mathbf{\Lambda_G}) = wR^A + (1-w)\sum_{g=1}^{G} R_g^{SS} \quad (6)$$

where R^A is an agreement criterion for all groups, R_g^{SS} a simple structure rotation criterion for group g, and $w \in [0, 1]$ a weight to assign relative importance to these two criteria. When optimizing this combined multi-group rotation criterion (by means of constrained maximum likelihood estimation), the group-specific factor variances and covariances are allowed to differ across groups, which helps to unravel differences in loadings from differences in factor (co-)variances (De Roover & Vermunt, 2019). In this, MGFR is similar to AESEM in the sense that it rotates and rescales the parameters. However, as shown in Equation (6), it performs these two steps at the same time—so that both rotation and rescaling optimize the agreement—while not considering the item intercepts. This makes MGFR the better alternative when the focus lies on investigating metric MI.

R_g^{SS} can currently be oblimin or geomin, or a target rotation toward an assumed measurement model.[3] For R^A, De Roover and Vermunt (2019) present two possible choices, namely, *generalized procrustes* (GP; Ten Berge, 1977) and *loading alignment* (LA). GP minimizes large loading differences between groups while allowing small differences. This is achieved by applying the least squares principle:

$$R_{GP}^A = \sum_{g_k=1}^{G}\sum_{g_l=g_k+1}^{G}\sum_{p}\sum_{m}(\lambda_{g_k pm} - \lambda_{g_l pm})^2 \quad (7)$$

Here, $\lambda_{g_k pm}$ is the loading of item p on factor m in group g_k. Although GP is originally an orthogonal rotation, the solution can be oblique because MGFR combines it with an oblique simple structure rotation.

LA is closely related to the alignment function in Equation (5) but considers only the loadings:

$$R_{LA}^A = \sum_{g_k=1}^{G}\sum_{g_l=g_k+1}^{G}\sum_{p}\sum_{m}\sqrt{\sqrt{(\lambda_{g_k pm} - \lambda_{g_l pm})^2 + \epsilon}} \quad (8)$$

where ϵ is again a small number to ensure continuous differentiability. LA pushes small loading differences to 0 while allowing (few) large differences. Because loading differences are then either 0 or large, LA is suitable to disentangle non-invariant and invariant loadings. Despite this theoretical advantage, MGFR with GP as the R^A criterion performed much better in the simulation studies by De Roover and Vermunt (2019).

All EFA-based MI-methods inherit the challenge of resolving rotational indeterminacy in the multi-group case. In the following empirical demonstration of the methods, we thus show how they can be combined with MGFR to achieve interpretable and comparable factor solutions. Table 1 provides an overview of the assumptions, hyperparameters, and capabilities of the presented EFA-based MI methods.

[3]Varimax rotation is also available but—in most cases—less ideal because it does not allow to disentangle differences in factor loadings from differences in factor (co-)variances.

7. Empirical Demonstration

7.1. Data

For our empirical demonstration of the presented methods, we used the dataset published by Bago et al. (2022) and investigated MI of the *Oxford Utilitarianism Scale* (OUS; Kahane et al., 2018). In a multilab study, Bago et al. (2022) examined the influence of psychological and situational factors on the judgement of moral dilemmas. Following Bago et al. (2022), we excluded participants who showed patterns of careless responding (i.e., wrong answers to control questions), indicated to have had technical problems, and did not answer the material in their native language. For simplicity of the subsequent analyses, we deleted all rows that contained missing values. This led to a final sample size of $N = 21746$.

The OUS measures utilitarian thinking, that is, how strongly people believe that actions should always aim at maximizing the overall good. It consists of two independent subscales, *impartial beneficence* (IB; measured by five items) and *instrumental harm* (IH; measured by four items). IB describes the attitude that no individual is more important than another, while IH means that moral rules can be neglected if it is for a greater good. Participants indicated their agreement to the items on a seven-point Likert scale (1 = "strongly disagree", 4 = "neither agree nor disagree", 7 = "strongly agree"). The items of the OUS can be found in the Appendix. We refer interested readers to Kahane et al. (2018) for more details on the OUS. Although assumptions about which items belong to which subscale are available, we only considered EFA models in our empirical demonstration (i.e., all items are allowed to load on both factors IB and IH). This let us illustrate in more detail one of the advantages of EFA-based MI investigations: they yield a more detailed picture of loading non-invariance by also taking into account (differences in) cross-loadings.

The data further contain many covariates for which (a violation of) MI of the OUS can be investigated. We did not consider every covariate with every method. Rather, we selected for each method the covariate(s) that we think best demonstrate(s) the main advantages of the respective method. In general, we applied all methods simply for didactic purposes to showcase their exemplary application. Our recommendation is *not* to always apply all methods. We looked at the following covariates:

- level of religiosity: continuous on a scale from 1 (lowest) to 10 (highest); $M = 4.21$ $SD = 2.79$
- region: categorical with three levels "Southern" ($N = 4692$), "Eastern" ($N = 2762$), and "Western" ($N = 14292$)

Table 1. Overview of methods based on exploratory factor analysis.

Method	Assumptions	Hyperparameters	Covariates	Level of MI tested	Result
MG-EFA (Dolan et al., 2009) Available in: R, Mplus, Latent Gold	Observed covariates (e.g., region) define potentially non-invariant groups (e.g., eastern and western region)	Level of significance for hypotheses tests	• Only categorical covariates • Only covariates with limited number of groups	All levels (configural, metric, scalar, residual)	• Group-specific parameter estimates • Hypothesis tests and fit indices for different levels of MI
MMG-EFA (De Roover, 2021; De Roover et al., 2022) Available in: R, Latent Gold	• There are clusters of groups for which parameters are invariant • For now: data are multivariate normally distributed	Number of clusters to be extracted	• Only categorical covariates • Covariates can have many groups	Metric and/or scalar (between unobserved clusters)	• Unobserved clusters of groups • Sets of cluster-specific, invariant parameter estimates
AESEM (Asparouhov & Muthén, 2023) Available in: Mplus	Most parameters are invariant, only few parameters are non-invariant	• Level of significance for hypotheses tests • Additive constant in loss function to ensure differentiability	• Only categorical covariates • Covariates can have many groups	No test of a specific level of MI (tests parameter difference for every parameter across all groups)	• Set of invariant parameters (or indication for which specific groups the parameter is not invariant) • Aligned factor scores that can be compared across groups
EFA trees (Sterner & Goretzko, 2023) Available in: R	For continuous covariates (e.g., age): there are two discrete groups defined along the covariate for which MI is violated (i.e., there are no gradual parameter differences)	• Level of significance for hypotheses tests • Maximum depth of trees • Minimum sample size in each leaf node	Both categorical and continuous covariates	• No test of a specific level of MI (tests parameter instability across all covariates) • Cannot find differences in intercepts (scalar MI)	• Groups defined by (interactions between) covariates across which MI is violated • Group-specific parameter estimates
MGFR (De Roover & Vermunt, 2019) Available in: Latent Gold	• For GP agreement: More large differences in loadings, less small differences • For LA agreement: Less large differences in loadings, more small differences	• Rotation criterion • Agreement criterion • Weight between rotation and agreement	• Only categorical covariates • Number of groups (= number of resulting loading matrices) should still be comparable and interpretable	Metric MI	• Group-specific loading matrices (rotated to simple structure and agreement) • Group-specific factor covariance matrices • Hypotheses tests for loading equivalence across groups

Note. MI = Measurement invariance, (M)MG = (Mixture) multi-group, EFA = Exploratory factor analysis, AESEM = Multi-group Exploratory Factor Alignment, MGFR = Multi-group factor rotation, GP = Generalized procrustes, LA = Loading alignment. Hyperparameters are parameters that cannot be estimated by data but have to be set prior to the analyses.

- age: continuous; $M = 26.05$ $SD = 10.25$
- gender: categorical with four levels "male" ($N = 6300$), "female" ($N = 15189$), "other" ($N = 63$), and "I wish not to answer" ($N = 194$)
- country of origin: categorical with 45 levels.

7.2. Software

The analyses were run in *R* (version 4.3.1; R Core Team, 2021), *Mplus* (version 8.9), and *Latent Gold* (Vermunt & Magidson, 2016), depending on which method is available in the respective software (see also Table 1). For analyses in R, we used the packages *lavaan* (Rosseel, 2012), *semTools* (Jorgensen et al., 2022), *partykit* (Hothorn & Zeileis, 2015), *mixmgfa* (available at https://github.com/KimDeRoover/mixmgfa/). Additionally, we created the R package *EFAtree* (https://github.com/philippsterner/EFAtree) which implements the EFA trees presented by Sterner and Goretzko (2023). The paper was written using the package *papaja* (Aust & Barth, 2020). All code needed to reproduce the analyses is openly available at https://osf.io/n8x5d/.

8. Results

8.1. MG-EFA

To demonstrate the use of MG-EFA, we investigated MI of the OUS on the covariate region, that is, between eastern, southern, and western participants. Table 2 shows that the configural model (with two latent factors for all groups) has an acceptable model fit. The χ^2-difference tests for both the comparisons of the configural and the metric as well as the metric and the scalar model is significant (both p-values are < 0.005). Judging by these test results, we would have to conclude that neither loadings nor intercepts are equal across the three groups. However, as mentioned, the χ^2-difference test is highly sensitive to sample size, which is quite large for the data set at hand. Differences in fit indices

Table 2. Results of multi-group exploratory factor analysis between regions.

Model	χ^2	df	*p*-value	RMSEA	Δ RMSEA	CFI	Δ CFI
Configural	1,185.15	57	0.000	0.052	0.000	0.958	0.000
Metric	1,339.90	85	0.000	0.045	−0.007	0.953	−0.005
Scalar	2,343.10	99	0.000	0.056	0.011	0.916	−0.037

Note. χ^2 = Value of the test statistic, df = Degrees of freedom, RMSEA = Root mean square error of approximation, Δ RMSEA = Difference in RMSEA between models, CFI = Comparative fit index, Δ CFI = Difference in CFI between models. A *p*-value of 0 means that it is < 0.001.

reveal that the fit of the metric model (where loadings are constrained to be equal across groups) is not worse than the fit of the configural model (ΔRMSEA = −0.01, ΔCFI = 0.00). Based on cut-off criteria for a comparison of a small number of groups (in our case: three), the conclusion that metric MI is supported seems more suitable (Chen, 2007; Cheung & Rensvold, 2002). When additionally constraining intercepts to be equal across groups, the fit becomes worse (ΔRMSEA = 0.01, ΔCFI = −0.04). Consequently, scalar MI seems to not be supported. We could conclude that latent covariances or relations (e.g., the correlation between IB and IH) can be compared between the three groups. Latent factor means, on the other hand, should not be compared without additional considerations (e.g., before establishing partial scalar MI).

As mentioned, the primary focus of EFA-based methods is to investigate differences in both main- and cross-loadings between groups. Although for this specific sample metric MI seems to be supported (evaluated across all items), it might still be informative to investigate the loadings of the individual items between groups. This lets us identify problematic items that could be changed or dropped to increase the invariance of the total scale. To achieve loading matrices that are comparable across groups, we used MGFR with oblimin rotation for all groups and GP as the agreement criterion. The weight of the agreement criterion was set to 0.5, as recommended starting settings by De Roover and Vermunt (2019; for more detailed recommendations on how to set the weight *w*, see Figure 1 in De Roover & Vermunt, 2019). Additionally, MGFR as implemented in Latent Gold provides Wald hypothesis tests that indicate which loadings on which factor significantly differ across groups. Of course, Wald hypothesis tests to identify significant differences in loadings could also be applied with any other rotation method.

Table 3 shows the resulting loading matrices. Table 4 shows the results of Wald hypothesis tests of loading invariance across the three regions. Due to multiple testing, we corrected the *p*-values with the Benjamini-Hochberg correction to control the false-discovery rate (Benjamini & Hochberg, 1995), using a level of significance of 0.05 (in the table, the corrected *p*-values are reported). Many main- and cross-loadings are significantly non-invariant across the three regions. For items 1, 5, 6, and 8 on factor IH and item 1, 2, 3, 5, 6, 8, and 9 on factor IB, the null hypothesis of loading invariance is supported by the data. However, because of the large sample size, these hypothesis tests have a high power to detect even small (and possibly irrelevant) loading differences. It is thus important to also inspect the

Table 3. Unstandardized loading matrices of multi-group exploratory factor analysis of the Oxford utilitarianism Scale with region as grouping covariate.

	Eastern		Southern		Western	
Items	IH	IB	IH	IB	IH	IB
Item 1	0.30	0.74	0.29	0.81	0.26	0.77
Item 3	0.01	1.20	0.11	1.20	0.16	1.17
Item 5	−0.17	0.78	−0.25	0.88	−0.26	0.88
Item 7	0.11	0.67	0.20	0.50	−0.02	0.69
Item 9	0.02	1.02	−0.11	0.94	−0.03	0.91
Item 2	1.00	0.16	1.12	0.13	1.14	0.15
Item 4	0.75	−0.05	0.51	0.20	0.58	0.10
Item 6	1.03	−0.04	0.97	−0.08	1.00	−0.06
Item 8	1.14	0.03	1.16	−0.03	1.11	−0.03

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. The factor solutions were obtained by applying multi-group factor rotation with oblimin rotation for all groups and generalized procrustes as the agreement criterion. The weight of the agreement criterion was set to 0.5.

Table 4. Results of Wald hypothesis tests of loading invariance across the three regions after multi-group exploratory factor analysis.

Factor	Item	Test statistic	df	p-value
IH	Item 1	1.99	2	0.370
	Item 2	17.01	2	0.000
	Item 3	24.89	2	0.000
	Item 4	24.48	2	0.000
	Item 5	4.67	2	0.146
	Item 6	2.37	2	0.349
	Item 7	37.91	2	0.000
	Item 8	3.65	2	0.206
	Item 9	12.16	2	0.004
IB	Item 1	2.52	2	0.420
	Item 2	0.80	2	0.670
	Item 3	1.18	2	0.664
	Item 4	27.86	2	0.000
	Item 5	6.31	2	0.097
	Item 6	1.05	2	0.664
	Item 7	25.88	2	0.000
	Item 8	4.48	2	0.198
	Item 9	7.00	2	0.090

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. df = Degrees of freedom. A combination of item and factor indicates for which item the invariance of loadings of this item on which factor was tested. For example: Item 1 and IB shows result of test of invariance of loadings of item 1 on factor IB across the three regions. *p*-values are Benjamini-Hochberg corrected. A *p*-value of 0.000 indicates that it is < 0.001.

loading matrices to pinpoint especially critical items.[4] Most notable are the loading differences between regions on items 7 ("It is just as wrong to fail to help someone as it is to actively harm them yourself.") and 4 ("If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used."). The main loading of item 7 is lower for the southern region (compared to eastern and western regions), while it also has a higher cross-loading in this group. Similarly, on item 4, both the southern and the western region have a lower

[4]Effect sizes for MI are available that are independent of the sample size, for example *EPC-interest* (Oberski, 2014), d_{MACS} (Nye & Drasgow, 2011), and extensions of d_{MACS} (Gunn et al., 2020). However, these effect sizes are not (yet) applicable to models with cross-loadings. To identify critical items, it might thus be advisable to inspect items with the highest differences in loadings across groups relative to each other. As an outlook, researchers might use the outlying-variable detection method proposed by De Roover et al. (2017), which is also applicable to loadings of an EFA.

Table 5. Fit statistics for the ten mixture multi-group exploratory factor analyses of the Oxford utilitarianism Scale.

Number of clusters	log L	fp	BIC	CHull scree ratio
1	−348,941.3	707	700,354.7	NA
2	−348,826.6	722	700,177.7	1.54
3	−348,752.3	737	700,081.6	1.63
4	−348,706.8	752	700,042.9	1.32
5	−348,672.3	767	700,026.3	1.14
6	−348,641.9	782	700,018.0	1.74
7	−348,624.4	797	700,035.5	1.17
8	−348,609.8	812	700,058.8	NA
9	−348,594.5	827	700,080.5	1.21
10	−348,582.1	842	700,108.2	NA

Note. log L = loglikelihood, fp = number of free parameters, BIC = Bayesian information criterion, CHull = convex hull. NAs can sometimes occur in the CHull procedure. Raising the number of random starts might alleviate this issue but in our case, even with 100 random starts some solutions fell under the hull.

main-loading than the eastern region, where the southern region again shows a notable cross-loading of 0.20 on this item. Attempts to increase MI of the OUS between regions could start with these two items.

8.2. MMG-EFA

We used MMG-EFA to unravel loading non-invariance of the OUS with regard to the covariate country. MMG-EFA is especially useful for this covariate because there are a large number of different countries, specifically 45 countries, in the data. While it is very unlikely that they all share the same loadings, it is plausible to assume that there are clusters of countries for which loadings are invariant. To allow for reliable estimations in each potential cluster, we only considered countries with sample sizes larger than 200. This led to 33 countries being considered in the analysis.

First, we conducted MMG-EFAs with one to ten clusters. According to both the BIC (with the number of groups as sample size) and the CHull procedure, the suggested number of clusters is six (see Table 5), which is the solution we selected.

Table 6 shows the composition of these six clusters. Each country was assigned to the cluster for which its posterior cluster membership probability $\hat{z}_{gk}$ was highest. It should be mentioned that $\hat{z}_{gk}$ can take on any value between zero and one, which allows groups to have high posterior cluster membership probabilities for more than one cluster. In practice, however, classification uncertainty is rare and limited because groups usually contain enough sample size for the model to be quite certain about their classification. In addition to the names of the countries, we added the region (as assigned in Bago et al., 2022) as well as the geographic region (i.e., the continent a country belongs to). The clustering does not seem to follow an obvious structure in terms of regions or continents. However, what can be concluded is that the loadings within each cluster are invariant (given that the correct number of clusters was selected). Two countries, China and Hungary, have their own cluster, which means that they do not share equivalent loadings with any other country.[5]

[5]If a group has its own cluster, it is important to check whether the sample size is large enough to allow for reliable estimations in this cluster. For our example, this was the case ($n_{China} = 1,175$ and $n_{Hungary} = 863$).

Table 7 shows the loading matrix of each cluster after a MGFR (i.e., the cluster membership was used as a new grouping covariate). Again, the weight of the GP agreement criterion was set to 0.5. Table 8 shows the results of Wald hypothesis tests of loading invariance across the six clusters for all combinations of items and factors. As can be seen, all main- and cross-loadings but one are significantly non-invariant (after Benjamini-Hochberg correction). Only for item 5 on the factor IH (which is a cross-loading), the null hypothesis of loading invariance is supported by the data. When comparing the loading matrices across clusters, we can see that for some items in some clusters there was a shift in main- and cross-loadings (e.g., item 1 in cluster 3 and 5, or item 4 in cluster 1). Additionally, some items show large cross-loadings in some clusters, whereas there are no cross-loadings on these items in other clusters. For example, item 8 has a large cross-loading in cluster 3 and 6, but no cross-loading in cluster 1 and 2.

A few things should be noted here: First, the large sample size leads to a high power of the Wald hypothesis test, rendering even practically irrelevant loading differences between clusters statistically significant. Second, the higher the number of clusters (or groups, in general), the more difficult it becomes for MGFR to rotate the loading matrices to a solution that is interpretable within each cluster but also comparable between clusters. Thus, it might be beneficial to change the rotation or agreement criterion as well as try different weights between these two criteria according to recommendations by De Roover and Vermunt (2019). This might yield results that are easier to interpret. Table 9 shows the loading matrices when the weight of the agreement criterion was set to 0.1 (i.e., putting more emphasis on simple structure rotation). In our case, while some loadings changed in size, the positions of main- and cross-loadings did not change notably.

In general, we could now inspect the item content and link loading differences between clusters (i.e., countries) to theory and empirical evidence. By doing so, we might rephrase items to increase the invariance of the OUS. Alternatively, we could also continue the analyses per cluster because loadings are invariant within each cluster. For example, we could investigate scalar MI within each cluster and simply refrain from comparing countries from different clusters.

8.3. EFA Trees

Because EFA trees can simultaneously evaluate multiple covariates for MI, we investigated MI with regard to level of religiosity, region, gender, and age. We used the following settings: level of significance was set to $\alpha = 0.005$, the maximum tree depth to three (including the first node, i.e., a maximum number of two splits), and the minimum sample size per node to $n = 400$. These settings allow for an interpretable tree but with possible interactions (due to restricted tree depth) and reliable parameter estimates in each node (due to large minimum sample size per node). At the same time, the low level of significance mitigates the risk of

Table 6. Composition of the clusters for the six-cluster solution of mixture multi-group exploratory factor analysis.

Cluster	Continent	Region	Country
Cluster 1	Americas	Southern	Argentina
	Americas	Western	Brazil
	Americas	Southern	Colombia
	Asia	Eastern	India
	Europe	Western	Italy
	Europe	Western	Netherlands
	Europe	Western	Portugal
	Europe	Western	Romania
	Asia	Southern	Turkey
Cluster 2	Oceania	Western	Australia
	Europe	Western	Austria
	Europe	Western	Bulgaria
	Americas	Western	Canada
	Europe	Western	Switzerland
	Europe	Western	Germany
	Europe	Western	Denmark
	Europe	Western	Spain
	Europe	Western	Greece
	Europe	Western	Croatia
	Europe	Eastern	North Macedonia
	Asia	Eastern	Pakistan
	Asia	Southern	Philippines
	Europe	Western	Serbia
	Europe	Southern	Slovakia
	Americas	Western	United States of America
Cluster 3	Asia	Eastern	China
Cluster 4	Europe	Southern	Czechia
	Europe	Southern	France
	Europe	Western	United Kingdom of Great Britain and Northern Ireland
	Europe	Western	Poland
Cluster 5	Asia	Eastern	Japan
	Europe	Western	Russian Federation
Cluster 6	Europe	Southern	Hungary

Table 7. Unstandardized loading matrices of the mixture multi-group exploratory factor analysis of the Oxford utilitarianism Scale with clusters as grouping covariate.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
Items	IB	IH	IB	IH	IB	IH	IB	IH	IB	IH	IB	IH
Item 1	0.64	0.40	0.92	0.35	0.34	0.73	0.95	0.35	0.08	0.46	1.12	0.38
Item 3	1.21	0.22	1.24	0.26	0.90	0.31	1.05	0.40	0.94	0.09	1.24	0.32
Item 5	0.83	−0.07	0.77	−0.16	0.61	0.01	0.82	−0.10	0.93	−0.10	0.49	−0.04
Item 7	0.57	0.24	0.56	0.10	0.97	0.02	0.85	−0.03	0.99	0.04	0.62	0.08
Item 9	1.02	0.06	0.82	0.08	1.07	0.12	0.84	0.02	0.98	0.28	0.60	−0.07
Item 2	0.19	1.17	0.32	1.12	0.51	0.80	0.31	1.02	0.30	1.00	0.24	0.92
Item 4	0.41	0.37	0.18	0.58	0.06	0.76	0.10	0.78	0.33	0.64	0.04	1.00
Item 6	0.28	0.81	0.04	0.97	0.03	1.12	−0.01	1.09	0.25	0.88	−0.06	1.04
Item 8	0.07	1.21	0.08	1.09	0.49	0.88	0.23	0.91	0.16	1.21	0.34	0.64

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. The factor solutions were obtained by applying multi-group factor rotation with oblimin rotation for all groups and generalized procrustes as the agreement criterion. The weight of the agreement criterion was set to 0.5.

Table 8. Results of Wald hypothesis tests of loading invariance across the six clusters of mixture multi-group exploratory factor analysis.

Factor	Item	Test statistic	df	p-value
IH	Item 1	61.22	5	0.000
	Item 2	76.76	5	0.000
	Item 3	50.23	5	0.000
	Item 4	135.13	5	0.000
	Item 5	14.99	5	0.010
	Item 6	65.57	5	0.000
	Item 7	46.32	5	0.000
	Item 8	120.13	5	0.000
	Item 9	27.69	5	0.000
IB	Item 1	323.90	5	0.000
	Item 2	61.38	5	0.000
	Item 3	74.39	5	0.000
	Item 4	77.17	5	0.000
	Item 5	38.38	5	0.000
	Item 6	102.41	5	0.000
	Item 7	111.62	5	0.000
	Item 8	128.33	5	0.000
	Item 9	74.06	5	0.000

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. df = Degrees of freedom. A combination of item and factor indicates for which item the invariance of loadings of this item on which factor was tested. For example: Item 1 and IB shows result of test of invariance of loadings of item 1 on factor IB across the four clusters. *p*-values are Benjamini-Hochberg corrected. A *p*-value of 0.000 indicates that it is < 0.001.

Table 9. Unstandardized loading matrices of the mixture multi-group exploratory factor analysis of the Oxford utilitarianism Scale with clusters as grouping covariate with more weight on rotation than on agreement.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5		Cluster 6	
Items	IB	IH	IB	IH	IB	IH	IB	IH	IB	IH	IB	IH
Item 1	0.62	0.35	0.91	0.24	0.24	0.71	0.92	0.24	0.02	0.47	1.13	0.19
Item 3	1.22	0.10	1.23	0.11	0.86	0.23	1.02	0.27	0.97	0.03	1.25	0.10
Item 5	0.86	−0.17	0.81	−0.26	0.62	−0.04	0.84	−0.20	0.98	−0.16	0.51	−0.13
Item 7	0.56	0.18	0.57	0.03	0.97	−0.07	0.86	−0.14	1.03	−0.03	0.63	−0.03
Item 9	1.04	−0.05	0.82	−0.02	1.06	0.03	0.84	−0.08	0.98	0.22	0.62	−0.18
Item 2	0.07	1.20	0.20	1.12	0.40	0.77	0.21	0.99	0.16	1.01	0.19	0.90
Item 4	0.38	0.34	0.12	0.57	−0.03	0.76	0.02	0.78	0.25	0.64	−0.02	1.02
Item 6	0.21	0.81	−0.07	1.00	−0.12	1.13	−0.12	1.10	0.13	0.89	−0.12	1.08
Item 8	−0.06	1.25	−0.04	1.11	0.38	0.84	0.14	0.90	−0.01	1.24	0.31	0.59

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. The factor solutions were obtained by applying multi-group factor rotation with oblimin rotation for all groups and generalized procrustes as the agreement criterion. The weight of the agreement criterion was set to 0.1.

Table 10. Hypothesis test result in the parent node of the EFA tree.

	Region	Gender	Religiosity	Age
Test statistic	577.966	595.155	294.987	452.478
p-value	0.000	0.000	0.000	0.000

Note. Test statistics were a χ^2 statistic for categorical and a supLM statistic for continuous covariates. A *p*-value of 0.000 indicates that it is < 0.001. If multiple *p*-values are below the level of significance, the covariate with the smallest *p*-value is selected.

finding practically irrelevant but statistically significant non-invariance. Note that we can only "afford" these rather strict settings because of the large sample. EFA trees could also be used with more liberal settings in smaller samples (see Sterner & Goretzko, 2023).

Table 10 shows the hypothesis test results in the parent node of the EFA tree. All *p*-values are below α, so the covariate with the smallest *p*-value is selected for splitting (in this case region with a *p*-value of 2.9×10^{-89}). The tree split the data into a group with only eastern observations and a group with both southern and western observations. Tables 11 and 12 show the hypothesis test results in these two resulting nodes, respectively. In both nodes, the EFA tree split the data on the covariate age[6], resulting in four final leaf nodes:

- eastern participants with age 27 or younger,
- eastern participants with age 28 or older,
- southern or western participants with age 24 or younger,
- southern or western participants with age 25 or older.

Figure 1 illustrates this tree structure with corresponding sample sizes in the leaf nodes. It is very likely that the tree would have continued to split the data, had we allowed deeper trees. However, this would have decreased the interpretability because it might have led to eight leaf nodes. If interpretation of the resulting partitions (i.e., potential three-way interactions between covariates) provides substantial increase in information gained, deeper trees are easily possible (e.g., see Brandmaier et al., 2013 for a SEM tree with eight leaf nodes). We continued our investigation with four leaf nodes.

[6]It is a coincidence that the tree further split the data on the covariate age in both nodes. It might have also happened that another covariate for splitting the data was chosen in one (or both) nodes.

Table 11. Hypothesis test result in the Eastern node of the EFA tree.

	Region	Gender	Religiosity	Age
Test statistic	0.000	119.487	70.671	121.313
p-value	NA	0.005	0.001	0.000

Note. Test statistics were a χ^2 statistic for categorical and a supLM statistic for continuous covariates. A *p*-value of 0.000 indicates that it is < 0.001. If multiple *p*-values are below the level of significance, the covariate with the smallest *p*-value is selected. The covariate region was not tested in this node because with only eastern observations, no further split on the covariate region is possible.

Table 12. Hypothesis test result in the Southern and Western node of the EFA tree.

	Region	Gender	Religiosity	Age
Test statistic	251.377	568.532	218.897	454.176
p-value	0.000	0.000	0.000	0.000

Note. Test statistics were a χ^2 statistic for categorical and a supLM statistic for continuous covariates. A *p*-value of 0.000 indicates that it is < 0.001. If multiple *p*-values are below the level of significance, the covariate with the smallest *p*-value is selected.

EFA trees do not provide information on which parameters differ between these four groups. For now, we can only conclude that there is a violation of MI with regard to an interaction between the covariates region and age. To better understand possible sources of non-invariance, EFA trees can be combined with MGFR. The node membership can be treated as a new grouping covariate. By applying MGFR, the loading matrices in the nodes can be rotated to increase interpretability of the parameters within nodes while also ensuring comparability between nodes (i.e., groups).

Table 13 shows the loading matrices for all four-leaf nodes after applying MGFR (with the weight of the GP agreement criterion set to 0.5). Table 14 shows the results of Wald hypothesis tests of loading invariance across the four-leaf nodes for all combinations of items and factors. Again, many main- and cross-loadings are significantly non-invariant (after Benjamini-Hochberg correction). Only for items 1, 5, 6, 8, and 9 on factor IH and item 2, 6, 8, and 9 on factor IB, the null hypothesis of loading invariance is supported by the data. Item 7 ("It is just as wrong to fail to help someone as it is to actively harm them yourself.") sticks out in both "younger" leaf nodes. For younger eastern participants, item 7 shows a lower main-loading compared to both "older" leaf nodes and also has the highest cross-loading in this leaf node. For younger southern-western participants, it has the lowest

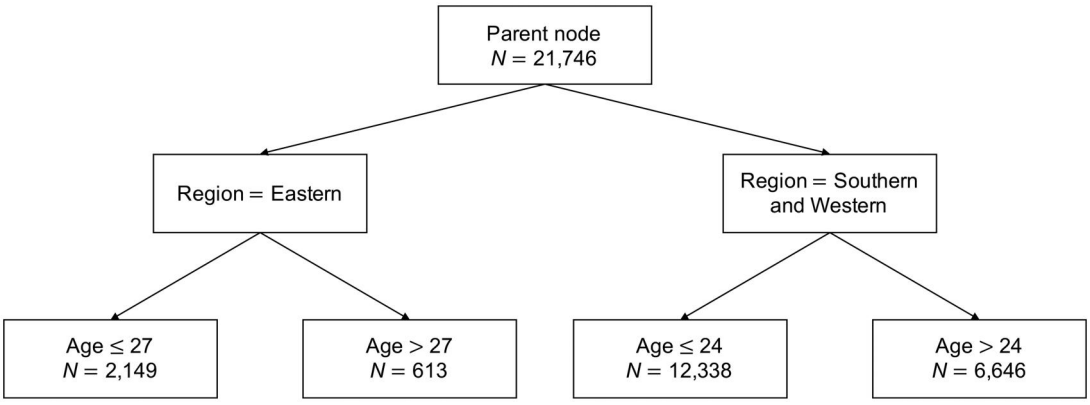

Figure 1. Resulting partition after applying EFA trees to the Oxford Utilitarianism Scale data.

Table 13. Unstandardized loading matrices of the exploratory factor analysis tree of the Oxford utilitarianism Scale with tree leaf nodes as grouping covariate.

	Eastern, Age ≤ 27		Eastern, Age > 27		South-West, Age ≤ 24		South-West, Age > 24	
Items	IH	IB	IH	IB	IH	IB	IH	IB
Item 1	0.32	0.75	0.46	0.58	0.33	0.80	0.31	0.67
Item 3	0.05	1.20	0.20	0.94	0.18	1.18	0.22	1.12
Item 5	−0.15	0.75	−0.05	0.77	−0.20	0.82	−0.24	0.94
Item 7	0.24	0.60	−0.12	0.99	0.08	0.56	0.14	0.71
Item 9	0.07	1.01	0.06	0.97	−0.01	0.91	−0.03	0.90
Item 2	1.03	0.17	0.96	0.20	1.19	0.14	1.18	0.14
Item 4	0.73	−0.07	0.79	0.12	0.52	0.15	0.65	0.14
Item 6	1.04	0.00	1.04	−0.11	1.00	−0.07	0.95	−0.02
Item 8	1.16	0.04	1.12	0.10	1.14	−0.02	1.14	−0.01

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. South-West stands for both regions southern and western. The factor solutions were obtained by applying multi-group factor rotation with oblimin rotation for all groups and generalized procrustes as the agreement criterion. The weight of the agreement criterion was set to 0.5.

Table 14. Results of Wald hypothesis tests of loading invariance across the four leaf nodes of the exploratory factor analysis tree.

Factor	Item	Test statistic	df	p-value
IH	Item 1	4.63	3	0.225
	Item 2	30.14	3	0.000
	Item 3	21.96	3	0.000
	Item 4	37.48	3	0.000
	Item 5	10.59	3	0.025
	Item 6	6.88	3	0.113
	Item 7	31.72	3	0.000
	Item 8	0.27	3	0.960
	Item 9	6.55	3	0.113
IB	Item 1	23.72	3	0.000
	Item 2	1.75	3	0.630
	Item 3	12.66	3	0.010
	Item 4	23.78	3	0.000
	Item 5	26.08	3	0.000
	Item 6	6.39	3	0.106
	Item 7	43.92	3	0.000
	Item 8	8.00	3	0.069
	Item 9	6.52	3	0.106

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively. df = Degrees of freedom. A combination of item and factor indicates for which item the invariance of loadings of this item on which factor was tested. For example: Item 1 and IB shows result of test of invariance of loadings of item 1 on factor IB across the four leaf nodes. *p*-values are Benjamini-Hochberg corrected. A *p*-value of 0.000 indicates that it is < 0.001.

cross- but also the lowest main-loading (in absolute terms). Item 3 has almost no cross-loading in the younger eastern node but has quite high cross-loadings (>0.20) in both older nodes (eastern and southern-western). Item 4 is especially noticeable in the younger southern-western node, where its main-loading is more than 0.10 lower compared to the older southern-western node and more than 0.20 lower than in both eastern nodes. Items 7 and 4 here too seem to be the most prominent items with regard to metric non-invariance.

8.4. AESEM

We used the four nodes generated by the EFA tree (cf. Figure 1) as groups across which AESEM is applied. AESEM could also be applied to a covariate with more groups (e.g., country with its 45 levels). However, by using the results of the EFA trees, we can keep the results easier to interpret while also demonstrating how the methods can be combined.

Table 15 shows the average loadings weighted by the sample size across all invariant groups (the detailed Mplus output is available at https://osf.io/n8x5d/). That is, these are the "most invariant" parameters that can be seen as estimates in the groups for which approximate MI holds. This is the case for almost all loadings. On the factor IB, only the loadings of item 7 (which is a main-loading) in the younger southern-western node, and of item 4 (cross-loading) in the younger eastern node were non-invariant compared to all other nodes. On the factor IH, the loadings of item 3 (cross-loading) in the eastern younger node, and item 4 (main-loading) in the younger southern-western node were non-invariant compared to all other nodes. Only for these two groups, these specific parameters cannot be seen as invariant estimates. It can be concluded that the proportion of non-invariant loadings is low; only four out of 72

Table 15. Unstandardized loading matrix of exploratory alignment of the Oxford utilitarianism Scale (weighted average loadings across invariant groups).

Items	IB	IH
Item 1	0.86	0.28
Item 3	1.32	0.11
Item 5	0.97	−0.26
Item 7	0.80	0.07
Item 9	1.04	−0.07
Item 2	0.14	1.19
Item 4	0.16	0.68
Item 6	−0.08	1.01
Item 8	−0.04	1.18

Note. IB and IH denote the latent factors impartial beneficence and instrumental harm, respectively.

loadings, that is, 5.6% are non-invariant (four nodes with 18 loadings each = 72 loadings). By applying AESEM to the nodes resulting from an EFA tree, we were thus able to achieve an approximately invariant set of loadings. This follow-up analysis could even be used to assess the reason why the EFA tree split the data. It might be that the EFA tree split the data because non-invariance on the four mentioned loadings was too large or because of other parameter differences (e.g., factor covariances or residual variances), given that most loadings are approximately invariant (after alignment).

8.5. Synthesis of the Results

The various analyses revealed some items that stick out with regard to non-invariance. Notably, items 4 and 7 showed lower main-loadings and higher cross-loadings compared to other items, both across regions (MG-EFA) and regions interacting with age (EFA trees and AESEM). Additionally, the analysis with MMG-EFA revealed that the investigation of MI across regions as defined by Bago et al. (2022) might not be too useful because many countries from the same region were assigned to different clusters. Taking these results together, it might be beneficial to investigate noticeable items against the background of the covariates across which they are non-invariant (here for example: items 4 and 7 across regions and age group). Additionally, results of the MMG-EFA can reveal potentially more adequate clusters of groups than what might be provided by prior classifications (e.g., the regions by Bago et al. (2022)).

In case of non-invariance, like in our example, the integration of the results also depends on the stage of the research process. Similar to single-group settings, EFA-based methods to investigate MI lend themselves to be applied already during scale development. In this stage, changes to the item pool are often still possible, and non-invariance could be addressed directly, for example by reformulating items 4 and 7 in the example above. Issues of non-invariance could then be prevented in the future. If a scale is already developed and EFA-based methods are applied to assess metric MI for both main- and cross-loadings, a violation of MI could also be seen as an interesting finding by itself. Using domain expertise, we could reason about potential causes of MI and model these causes accordingly or test our hypotheses about them (Sterner et al., 2024). If EFA-based methods are used as a precursor of CFA-based analyses, one could also aim for partial MI, for example by testing whether freeing certain main- or cross-loadings in a stepwise manner would improve the fit of the model. For this, the results of the EFA-based analysis, that is, which main- and cross-loadings were significantly different, could be taken into account, too.

In closing, we want to highlight again that we applied all methods for didactic purposes. The choice of methods for each individual application depends on the data set, the available covariates, the research question, and the assumptions one is willing to make (cf. Table 1).

9. Discussion

We presented EFA-based methods to investigate MI. The focus of these methods is on the investigation of metric MI, that is, the invariance of main- and cross-loadings across groups. For each method, we detailed the model specification as well as its advantages and drawbacks. We demonstrated the assumptions that have to be made and the insights we gain in return in an empirical example. On top of that, we showed how EFA-based MI methods can be combined with MGFR to resolve the rotational indeterminacy in multi-group settings.

The main take-away of our presentation and demonstration is that the optimal choice of a method depends on the question you want to answer, combined with the specificities of the data at hand. A detailed (yet not exhaustive) overview of prerequisites and capabilities of each method is given in Table 1. Ideally, the methods are combined to thoroughly scrutinize the data for MI. For example, the clusters or nodes resulting from MMG-EFA or EFA-trees, respectively, can be used as groups in MG-EFA, and the resulting loading matrices can be rotated by MGFR. In this, covariates with many groups can be reduced to a smaller number of clusters or nodes, for which we can then, for example, investigate scalar MI by means of hypothesis testing. We refrained from addressing scalar MI due to the focus of EFA-based methods on metric MI. However, as we will discuss shortly, EFA-based methods could function as a precursor of CFA-based investigations of scalar MI. Even further, clusters or nodes resulting from prior analyses could be used as groups in multi-group (E)SEM (Asparouhov & Muthén, 2009), allowing us to then model structural relations between our constructs of interest.

As mentioned in the introduction, EFA-based methods differ from CFA-based methods mainly in the fact that no (potentially overly restrictive) zero-loadings have to be imposed. This allows for a more detailed investigation of metric MI because violations of metric MI due to cross-loadings can be considered. However, the investigation of scalar MI is hampered. For example, with EFA trees, the intercepts cannot be included in the model estimation (in *lavaan* language: the argument *meanstructure* must not be set to *TRUE*). The intercepts of the items (the parameters we want to test for invariance) are intertwined with the factor means (the parameters we want to compare between

groups). Even if the intercepts were equal across groups, an EFA tree would split the data if the factor means were different between groups. Similarly, the results of MMG-EFA with clustering based on loadings do not consider (non-)invariance of intercepts. It is possible to cluster the groups on both loadings and intercepts. However, this entails the assumption that there is one underlying clustering for both of these sets of parameters (Leitgöb et al., 2023). It is thus more advisable to first cluster the groups based on the loadings and then, per obtained cluster, continue to cluster the groups based on the intercepts. When applying AESEM, both loadings and intercepts are considered and, thus, scalar invariance is also investigated. But because we are also investigating all cross-loadings for invariance, there are more loadings than intercepts that are being estimated and aligned (if the specified model has two or more factors). It should be examined whether this potential dominance of loadings (when minimizing Equation (5)) has some undesired effect on the assessment of scalar MI.

In summary, EFA-based methods are not to be seen as methods "competing" with CFA-based ones, for example the methods detailed in Kim et al. (2017). Rather, they are a useful addition to the MI-toolbox that broaden the capabilities of investigating MI along an exploratory-confirmatory continuum (Nájera et al., 2023). Especially in the context of scale development, it can be beneficial to apply EFA-based methods to investigate the violations of MI due to cross-loadings, before using CFA-based methods to assess scalar MI. An EFA-based method is able to identify potentially non-invariant cross-loadings and allows us to alter the model based on these results (which we then have to validate on new data, of course). This approach is superior to the aforementioned strategy of (repeatedly) modifying CFA models in a data-driven way because it capitalizes less on chance (MacCallum et al., 1992).

9.1. Future Research

All presented methods are rather new. While they have been investigated thoroughly in the original papers that introduced them, more simulated and empirical research is needed to better understand their behavior under various conditions. As mentioned in the introduction, the alignment method can be seen as the method that has been researched the most among all methods presented here (Flake & McCoach, 2018; Lomazzi, 2018; Luong & Flake, 2023; Munck et al., 2018), but much less so when cross-loadings are present (i.e., when AESEM is used). For all the methods at hand, not much is known about their performance when, for example, data are non-normal, covariates are highly correlated, or MI is violated in a nuanced way (for example, a U-shaped relation between values of a continuous covariate and parameter values).

More broadly, it would be interesting to research and demonstrate how exactly EFA-based methods can be used as a precursor of CFA-based analyses. On the one hand, this can be done in methodological studies, for example, by investigating the benefits of refining a model using EFA-based methods in the context of MI (e.g., instead of using modification indices in the context of CFA). On the other hand, and maybe even more importantly, tutorial papers are needed that showcase potential workflows of MI investigations to provide guidance for applied researchers. Somaraju et al. (2022) showed a workflow that details use cases and follow up analyses in the context of MG-CFA and alignment. Such workflows could be extended by preceding EFA-based analyses. In this, the different groups and models entering the CFA-based analyses would have been scrutinized, in an exploratory manner, for model (mis-)specifications and metric MI beforehand, hopefully allowing for a more well-founded investigation of scalar MI.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

References

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*, 397–438. https://doi.org/10.1080/10705510903008204

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

Asparouhov, T., & Muthén, B. O. (2023). Multiple group alignment for exploratory and structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 169–191. https://doi.org/10.1080/10705511.2022.2127100

Aust, F., & Barth, M. (2020). *Papaja: Create APA manuscripts with R Markdown*. https://github.com/crsh/papaja

Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albalooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., … Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*, 880–895. https://doi.org/10.1038/s41562-022-01319-5

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: Statistical Methodology*, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research*, *36*, 48–86. https://doi.org/10.1177/0049124107301947

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*, 139–158. https://doi.org/10.1177/1094428102005002001

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

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

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*, 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*, 492–514. https://doi.org/10.1198/106186008X319331

Brandmaier, A. M., Oertzen, T. v., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*, 71–86. https://doi.org/10.1037/a0030001

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Byrne, B. M., & Vijver, F. J. R. v. d. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*, 107–132. https://doi.org/10.1080/15305051003637306

Cao, C., & Liang, X. (2022a). Sensitivity of fit measures to lack of measurement invariance in exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 248–258. https://doi.org/10.1080/10705511.2021.1975287

Cao, C., & Liang, X. (2022b). The impact of model size on the sensitivity of fit measures in measurement invariance testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*, 744–754. https://doi.org/10.1080/10705511.2022.2056893

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. https://doi.org/10.1207/s15327906mbr0102_10

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

Ceulemans, E., & Van Mechelen, I. (2005). Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection. *Psychometrika*, *70*, 461–480. https://doi.org/10.1007/s11336-003-1067-3

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

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

De Roover, K. (2021). Finding clusters of groups with measurement invariance: unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*, 663–683. https://doi.org/10.1080/10705511.2020.1866577

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*, 216–229. https://doi.org/10.3758/s13428-015-0687-8

De Roover, K., & Vermunt, J. K. (2019). On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 905–923. https://doi.org/10.1080/10705511.2019.1590778

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

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*, 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

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*, 295–314. https://doi.org/10.1080/10705510902751416

Flake, J. K., & McCoach, D. B. (2018). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 56–70. https://doi.org/10.1080/10705511.2017.1374187

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*, 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*, 123–144. https://doi.org/10.1177/00131644231163813

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of six effect size measures of measurement non-invariance for continuous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*, 503–514. https://doi.org/10.1080/10705511.2019.1689507

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*, 3905–3909.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*, 36–44. https://doi.org/10.1038/s41562-019-0734-z

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, *38*, 593–604. https://doi.org/10.1007/BF02291497

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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*. https://CRAN.R-project.org/package=semTools

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131–164. https://doi.org/10.1037/rev0000093

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*, 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*, 491–515. https://doi.org/10.1177/0022022106290474

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

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, Data, Analyses: A Journal for Quantitative Methods and Survey Methodology (MDA)*, *12*, 77–103. https://doi.org/10.12758/mda.2017.09

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. IAP.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research*, *46*, 340–364. https://doi.org/10.1080/00273171.2011.564527

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial

for transparent analysis planning and reporting. *Psychological Methods*, *28*, 905–924. https://doi.org/10.1037/met0000441

Maassen, E., D'Urso, E. D., van Assen, M. A. L. M., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000624

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*, 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

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

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

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*, 111–130. https://doi.org/10.21500/20112084.857

Mulaik, S. A. (2010). *Foundations of factor analysis*. CRC Press.

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: The alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research*, *47*, 687–728. https://doi.org/10.1177/0049124117729691

Nájera, P., Abad, F. J., & Sorrel, M. A. (2023). Is exploratory factor analysis always to be preferred? A systematic comparison of factor analytic techniques throughout the confirmatory–exploratory continuum. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000579

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *The Journal of Applied Psychology*, *96*, 966–980. https://doi.org/10.1037/a0022955

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*, 45–60. https://doi.org/10.1093/pan/mpt014

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: DR*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial invariance in multiple populations: A multiple testing procedure. *Educational and Psychological Measurement*, *73*, 713–727. https://doi.org/10.1177/0013164412451978

Robitzsch, A. (2023). Implementation aspects in invariance alignment. *Stats*, *6*, 1160–1178. https://doi.org/10.3390/stats6040073

Robitzsch, A. (2022). *Sirt: Supplementary item response theory models*. https://cran.r-project.org/web/packages/sirt/index.html

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rudnev, M. (2019). *Alignment method for measurement invariance: Tutorial*. https://maksimrudnev.com/2019/05/01/alignment-tutorial/

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

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

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A review of measurement equivalence in organizational research: What's old, what's new, what's next? *Organizational Research Methods*, *25*, 741–785. https://doi.org/10.1177/10944281211056524

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Sterner, P., & Goretzko, D. (2023). Exploratory factor analysis trees: evaluating measurement invariance between multiple covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*, 871–886. https://doi.org/10.1080/10705511.2023.2188573

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A causal framework for the comparability of latent variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12. https://doi.org/10.1080/10705511.2024.2339396

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316. https://doi.org/10.1007/s11336-013-9388-3

Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, *42*, 267–276. https://doi.org/10.1007/BF02294053

Van Bork, R., Rhemtulla, M., Sijtsma, K., & Borsboom, D. (2022). A causal theory of error scores. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000521

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486–492. https://doi.org/10.1080/17405629.2012.686740

Appendix

Items and corresponding subscales of the OUS (Kahane et al., 2018).

ID	Item	subscale
1	If the only way to save another person's life during an emergency is to sacrifice one's own leg, then one is morally required to make this sacrifice.	IB
2	It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people.	IH
3	From a moral point of view, we should feel obliged to give one of our kidneys to a person with kidney failure since we don't need two kidneys to survive, but really only one to be healthy.	IB
4	If the only way to ensure the overall well-being and happiness of the people is through the use of political oppression for a short, limited period, then political oppression should be used.	IH
5	From a moral perspective, people should care about the well-being of all human beings on the planet equally; they should not favor the well-being of people who are especially close to them either physically or emotionally.	IB
6	It is permissible to torture an innocent person if this would be necessary to provide information to prevent a bomb going off that would kill hundreds of people.	IH
7	It is just as wrong to fail to help someone as it is to actively harm them yourself.	IB
8	Sometimes it is morally necessary for innocent people to die as collateral damage—if more people are saved overall.	IH
9	It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal.	IB

Note. IB = Impartial Beneficence; IH = Instrumental Harm.