

Department of Mathematics

---

Master Thesis

Winter 2021/2022

---

Pit Rieger

**Measurement Invariance in  
Confirmatory Factor Analysis:  
Methods for Detecting Non-invariant Items**

---

November 9, 2021

---

Adviser      Dr. Markus Kalisch  
Co-Adviser   Prof. Dr. Marco Steenbergen



## Abstract

*Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.*

# Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
1.1	Notation . . . . .	4
<b>2</b>	<b>Confirmatory Factor Analysis</b>	<b>4</b>
2.1	Refresher: Exploratory Factor Analysis . . . . .	6
2.2	Confirmatory Factor Analysis . . . . .	10
<b>3</b>	<b>Measurement Invariance</b>	<b>15</b>
3.1	Types of Measurement Invariance . . . . .	17
3.2	Partial Measurement Invariance . . . . .	19
3.3	Global Test of Measurement Invariance . . . . .	20
<b>4</b>	<b>Detecting Non-invariant Items</b>	<b>22</b>
4.1	Existing Methods for Detecting Non-invariant Items . . . . .	23
4.2	A Novel Approach to Non-invariant Item Detection . . . . .	27
4.3	Overview over Methods for Detecting Non-invariant items . . . . .	29
<b>5</b>	<b>Simulation Study</b>	<b>29</b>
5.1	Results . . . . .	32
<b>6</b>	<b>Application: Studying the Cross-National measurement invariance of Populism Scales</b>	<b>35</b>
6.1	Populism Scales . . . . .	35
<b>7</b>	<b>Conclusion</b>	<b>35</b>
<b>8</b>	<b>Appendix</b>	<b>40</b>
8.1	Derivation of the Log-Likelihood of the EFA Model . . . . .	40
8.2	Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI) . . . . .	41
8.3	Scale invariance of the EFA model . . . . .	41
<b>9</b>	<b>TEST SECTION - TO BE DELETED</b>	<b>41</b>

Indicator	$Y$
Latent variable	$\eta$

**Table 1:** Caption

## 1 Introduction

measurement invariance (MI, also referred to as measurement equivalence). In the framework of the closely related item response theory (IRT), this issue is commonly referred to as differential test functioning (DTF) or as differential item functioning (DIF) at the level of individual items (e.g. Drasgow et al., 2018; Thissen et al., 1993). Since the new method for detecting measurement non-invariance at the item level that is introduced in this thesis is not easily implemented in IRT, the remainder will be within the CFA framework. However, some scholars have attempted to unify the research on MI, DTF, and DIF (e.g. Stark, Chernyshenko, Drasgow, & Williams, 2006).

CFA was first introduced by Jöreskog (1969).

This thesis progresses as follows. First, a review of the basic concepts of the widely-known exploratory factor analysis (EFA) model is given before the confirmatory factor analysis (CFA) model is introduced. [ALSO REFERRED TO IN SEM LITERATURE AS Measurement MOdel] Next, the concept of measurement invariance is defined and it is shown how the problem of measurement noninvariance can arise in the CFA framework as well as the implications of noninvariance. After having established a theoretical foundation, a detailed review of the literature on measurement invariance and specifically detection of noninvariant items is provided. Then, a new method detecting noninvariant items is introduced and its performance evaluated with simulation data. Before turning to a final discussion and conclusion, an application to real-world data from the study of populist attitudes in cross-national political science research is presented.

### 1.1 Notation

## 2 Confirmatory Factor Analysis

Many scientific concepts that are of interest for social scientists cannot be observed directly. Quantitative researchers usually refer to such variables as being *latent*. While observable variables can simply be measured, latent variables have to be inferred with the help of a measurement model. A popular approach is confirmatory factor analysis (CFA) which constitutes a framework of several measurement models. In a nutshell, all CFA models circumvent the key obstacle by using several observable *indicators* that relate to the unobservable latent variable(s).

This section starts off with a motivating example to illustrate the need for measurement models such as CFA in the social sciences and to highlight some of the informal impli-

cations of using CFA when inferring latent variables. Before going into the formal introduction of the CFA approach, the well-known and very similar exploratory factor analysis (EFA) model is recapitulated. This makes it much easier to then elaborate on the subtle differences between the EFA and CFA model. For both methods, I introduce the formal setup along with the key assumptions, give some intuition as to how they can be estimated, as well as illustrate their idiosyncracies. By the end of this section, the reader should be well prepared to comprehend the problem of measurement non-invariance which will be discussed in the next section.

As a motivating example, suppose a group of researchers would like to study political ideology in Switzerland. In many European countries, one important part of people's political belief system can be summarized by their position on an (economic) left-right dimension. Put crudely, left-leaning citizens value social equality highly, which often entails support for redistributive policies and greater state intervention. Right-leaning citizens, on the other hand, are less concerned with the existence of social hierarchical structures and inequalities, which is often accompanied by a favorable position towards free-market solutions and a small state. Political ideology is a classic example of a latent construct because it can obviously not be observed directly. Many studies solve this issue by asking survey respondents to place themselves on a left-right scale, leaving analysts to make sense of their results and forcing them to take them at face value, which is problematic for several reasons. To name but a few, people may not be aware of their own position, they may be unfamiliar with the concept entirely, or they may have a different definition of its meaning than the researchers. To improve on this rather crude approach, CFA can be used to infer respondents' left-right position. This requires researchers to devise a battery of questions that are indicative of the latent construct, but leave less room for interpretation of the question. For example, items in the battery could ask respondents about their attitudes towards minimum-wage laws, free-trade agreements, or unionization. Researchers can then specify in a CFA model how these indicators relate to the latent variable and to one another. Using responses to the questions, they can then ultimately infer respondents' ideological positions.

The motivating example also highlights several non-technical fundamental implications of using CFA models. These will not play a large role in the formal introduction below owing to the fact that the specification of CFA models is typically done on the basis of expert knowledge and is highly field-dependent. Notwithstanding, model decisions regarding both the model specification and the choice/construction of indicators are highly consequential for the interpretation of the inferred latent variables. First, the construction and choice of a set of indicators influences what constitutes the inferred latent construct. This is a result of the simple fact that no single item will capture the full breadth of the latent construct and, vice versa, the latent construct will not be the only factor explaining variation in responses to the items. It is therefore crucial to construct the indicators in a way that reflects and covers the entire concept. Second, [in a similar spirit, the specification of a CFA model influences wha Second, for CFA measurement models that truly relate to latent

variables, the validity of the model can never be directly established. At best, A core obstacle is therefore to identify how and to what degree the indicators are related. Also note how crucial the concrete specification of the items is for any subsequent analysis. It defines what constitutes the fundamental components of the latent construct and thus what is actually being measured.]

## 2.1 Refresher: Exploratory Factor Analysis

### 2.1.1 Model Setup

The EFA setup supposes we have a set of  $p$  indicators  $\mathbf{Y} = (Y_1, \dots, Y_p)^\top$ , also referred to as *manifest variables* or *items*, that are assumed to relate to  $k$  latent variables  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^\top$  following some distribution with  $\mathbb{E}[\boldsymbol{\eta}] = \boldsymbol{\mu}$  and  $\text{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Phi}$ . For now, we assume that the number of latent variables  $k$  is known. However, this is typically not the case and we return to how we can choose  $k$  in section 2.1.5. We further assume the relationship to be of the following linear, multivariate, and multiple regression form:

$$\begin{aligned} Y_1 &= \tau_1 + \lambda_{11}\eta_1 + \dots + \lambda_{1k}\eta_k + \varepsilon_1 \\ &\vdots \\ Y_p &= \tau_p + \lambda_{p1}\eta_1 + \dots + \lambda_{pk}\eta_k + \varepsilon_p, \end{aligned} \tag{1}$$

where  $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)^\top$  are intercepts,  $\lambda_{ij}$  are regression coefficients, and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)^\top$  are errors. In the FA framework, we refer to the regression coefficients as (*factor*) *loadings* and the errors as *specific variables*, respectively. Note, that the assumed relationship implies that each indicator is a linear combination of all latent variables plus an intercept and an idiosyncratic error.

Writing the factor loadings as a  $p \times k$  loading matrix  $\boldsymbol{\Lambda}$ , we can equivalently write the model in matrix form as

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{2}$$

### 2.1.2 Assumptions

To emphasize,  $\boldsymbol{\Lambda}$  is unknown and  $\boldsymbol{\eta}$  is unobservable. As demonstrated by the motivating example, this is the fundamental reason for conducting factor analysis. However, in order to obtain estimates for these quantities, we require additional assumptions. With respect to the specific variables, we assume that they have mean zero, are pairwise uncorrelated,

and uncorrelated with the latent variables:

$$\mathbb{E}(\boldsymbol{\varepsilon}) = 0 \quad (3a)$$

$$\text{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}, \text{ a diagonal matrix} \quad (3b)$$

$$\text{Cov}(\boldsymbol{\eta}, \boldsymbol{\varepsilon}) = 0, \quad (3c)$$

These assumptions are fairly standard and resemble the usual assumptions of standard linear regression. Furthermore, they allow the decomposition of the covariance matrix of  $\mathbf{Y}$  as

$$\boldsymbol{\Sigma} := \text{Cov}(\mathbf{Y}) = \text{Cov}(\boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) \quad (4a)$$

$$\stackrel{(3c)}{=} \text{Cov}(\boldsymbol{\Lambda}\boldsymbol{\eta}) + \text{Cov}(\boldsymbol{\varepsilon}) \quad (4b)$$

$$= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}, \quad (4c)$$

where  $\boldsymbol{\Phi} := \text{Cov}(\boldsymbol{\eta})$ . It is clear from this decomposition that by assuming a model of the form in equation (2), we implicitly assume a covariance structure because  $\boldsymbol{\Sigma}$  clearly depends on the model parameters. Thus, it often makes sense to talk about model-implied covariances. We occasionally write  $\boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$  to emphasize this dependence. The model-implied covariance can then be used to establish the goodness-of-fit of the model by comparing it with the sample covariance matrix of the actual data. Taking things one step further, competing measurement models can then be compared by comparing their goodness-of-fit. Furthermore, the decomposition gives way to the ML estimation of FA models which will be discussed below.

### 2.1.3 Estimation

The EFA model is typically estimated by means of a maximum likelihood (ML) approach.<sup>1</sup> This, of course, requires an additional distributional assumption. Under multivariate normality, we can equivalently write the model as

$$\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Sigma}). \quad (5)$$

It is standard practice to work with centered manifest variables  $\mathbf{Y}' = \mathbf{Y} - \boldsymbol{\tau}$ , such that

$$\mathbf{Y}' \sim \mathcal{N}_p(\boldsymbol{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Sigma}), \quad (6)$$

---

<sup>1</sup> Another common alternative is the principal factor method. However, of the two, only ML estimation can be used for confirmatory factor analysis. I therefore only discuss the ML approach.

where in practice, we simply subtract the sample mean from each manifest variable to ensure their centering.

Estimates can then be obtained by maximizing the normal log-likelihood over the parameters in  $\Lambda$  and  $\Psi$ . Let  $\theta := (\Lambda, \Psi)$ , then the log-likelihood is given by

$$\ell(\Sigma(\theta) | \mathbf{S}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma(\theta)) - \frac{n}{2} \text{tr}(\mathbf{S}\Sigma^{-1}(\theta)), \quad (7)$$

where  $\mathbf{S}$  is the sample covariance matrix of  $\mathbf{Y}$ . A full derivation of the log-likelihood as well as its relationship with the Wishart distribution can be found in section 8.1 of the appendix. Note, however, that estimation has traditionally been conducted by equivalently minimizing the fit function

$$F(\Sigma(\theta) | \mathbf{S}) = \log \det(\Sigma(\theta)) - \log \det(\mathbf{S}) + \text{tr}(\mathbf{S}\Sigma^{-1}(\theta)) - p \propto \ell(\Sigma(\theta) | \mathbf{S}), \quad (8)$$

where the replacement of constants in equation (7) with  $\log \det(\mathbf{S})$  and  $p$  conveniently sets the fit function to zero when  $\Sigma(\theta) = \mathbf{S}$ .

#### 2.1.4 Rotational Invariance

It is important to observe a crucial obstacle for EFA and its estimation: *rotational invariance*. To see this, consider a matrix  $\mathbf{R}$  of dimension  $k \times k$  and let

$$\tilde{\Lambda} := \Lambda \mathbf{R} \quad (9a)$$

$$\tilde{\eta} := \mathbf{R}^{-1} \eta \quad (9b)$$

be transformed loadings and latent variables. Then their factor model is

$$\mathbf{Y} = \tilde{\Lambda} \tilde{\eta} + \varepsilon = \Lambda \mathbf{R} \mathbf{R}^{-1} \eta + \varepsilon = \Lambda \eta + \varepsilon \quad (10)$$

which, as the last identity shows, is equivalent to the model of the untransformed loadings and latent variables. In other words, the EFA model is only identifiable up to a simultaneous transformation of the loadings and latent variables so there is no unique solution for  $\Lambda$  and  $\eta$ . Note, that although  $\mathbf{R}$  is not strictly limited to rotation matrices, this property is called *rotational invariance*.

Due to rotational invariance, estimation of the EFA model is not possible without additional constraints. The standard solution is to impose constraints on the latent variables to render the model identifiable. These constraints amount to the latent variables having mean zero, unit variance, and being pairwise uncorrelated, i.e.



$$\mathbb{E}[\boldsymbol{\eta}] = 0 \quad (11a)$$

$$\text{Cov}(\boldsymbol{\eta}) = I_{k \times k}, \text{ an identity matrix.} \quad (11b)$$

Given these properties of  $\boldsymbol{\eta}$ , it is easy to see that  $\text{Cov}(\tilde{\boldsymbol{\eta}}) = I_{k \times k}$  if and only if  $\mathbf{R}$  is an identity matrix itself. The resulting unique solution under these constraints is therefore commonly referred to as the *unrotated solution*.

The unrotated solution can then still be subjected to post-estimation transformations  $\mathbf{R}$  while yielding model parameters that are equally valid because they only violate the arbitrary constraints on the latent variables. If  $\mathbf{R}$  is an orthogonal matrix, the transformation preserves the uncorrelatedness of the factors in the unrotated solution and we refer to it as an *orthogonal rotation*. Other transformations that are not orthogonal are called *oblique rotations*. Oftentimes, the goal of applying a rotation is to ease the interpretation of the factor loadings. In this regard, different algorithmic rotations have been proposed, which result in loading matrices that are more easily interpretable. For example, a prominent method, the orthogonal varimax rotation (Kaiser, 1958), tries to find a rotation such that each latent variable has few high and many vanishing loadings. Numerous other methods for obtaining rotated solutions exist (see Browne, 2001, for an overview). However, the supposed subjectivity involved in rotations has also been grounds for criticism (e.g. Horn, 1967; but also see Mulaik, 1987).

### 2.1.5 Choice of $k$

As mentioned above, in the exploratory setting in which EFA is mostly used, a “true” number of underlying latent variables  $k$  is typically unknown to the researcher or doesn’t even exist which is why  $k$  is often considered a tuning parameter. From a purely statistical point of view an EFA model can be estimated as long as the degrees of freedom are positive. The degrees of freedom  $d_k$ , given  $p$  manifest variables, are given by

$$d_k = \frac{(p - k)^2 - p - k}{2}. \quad (12)$$

Different methods for selecting the “optimal” number of factors, have been proposed (for overviews, see Preacher et al., 2013; Zwick & Velicer, 1986). In general, there exists a trade-off between the goodness-of-fit and a parsimonious model. The goal is to strike a balance by obtaining a parsimonious model with few latent variables that fits the data well. Most commonly, the choice of  $k$  is made on the basis of the scree test or scree plot (Cattell, 1966). Put briefly, EFA models are fit for all  $k$  for which they are still identified and the final  $k$  is then chosen in accordance with some rule of thumb, e.g. the share of variance in the sample covariance matrix that is explained by the model.

## 2.2 Confirmatory Factor Analysis

### 2.2.1 Model Setup

The fundamental setup for the CFA model remains almost identical to the EFA model. We still try to model our manifest variables as a linear function of latent variables. The core difference is the following: While EFA serves the purpose of data exploration and dimension reduction, CFA provides a way of including assumption about the structure of the model. By structure, I mean all aspects of the model that relate to relationships between any variable (both latent and manifest), their variances, arbitrary constraints on any parameter, the number of latent variables and many more aspects of the model. Recall that for the EFA model, we simply assumed that all latent variables load on all indicators and the only modelling choice is to set the number of latent variables and perhaps applying a rotation to the loadings. For a CFA model, we can for example model one latent variable to load on only a subset of indicator variables while another latent variable loads on a different (potentially overlapping) subset of indicator variables. In this sense, we can view the EFA model as a special case of the CFA. The great advantage of the flexibility of the CFA model is that it allows researchers to include knowledge from their area of expertise in the form of pre-specifications of the structure of the model. What's more, the CFA framework enables researchers to conduct statistical tests of their pre-specifications by treating them as hypotheses about the structure of the model and comparing their implications for the implied data generating process with observed data. Put differently, CFA treats these pre-specifications as a testable hypothesis about the structure of the data. How researchers arrive at these hypotheses is an important, if not the most important, aspect of CFA. That said, I assume throughout this thesis that the basic structure of the true model is known because the question of how to configure a CFA model from scratch is more a question of substance than statistics and varies heavily across domains.

Returning to the formal model, we can keep all prior notation and still write the model in matrix notation as

$$\mathbf{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \quad (13)$$

which formally implies the same model covariance matrix as the EFA model, given by

$$\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^T + \boldsymbol{\Psi}. \quad (14)$$

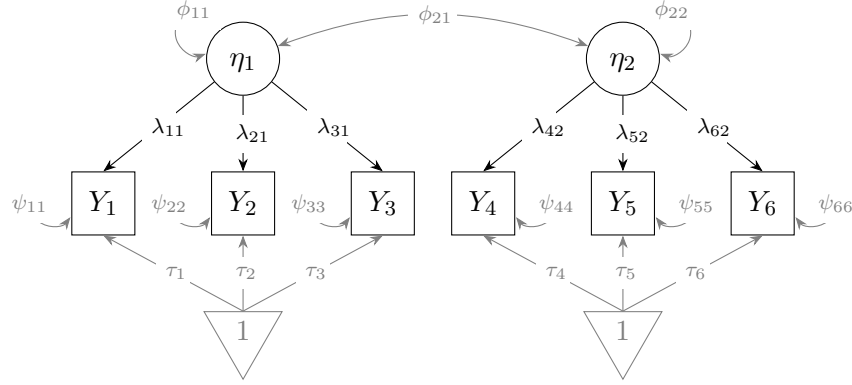
The somewhat hidden, yet decisive difference, is how we treat the parameters in this model. Broadly speaking, the pre-specifications about the structure of the model can be viewed as constraints on the parameters in the model. These constraints may relate to the intercepts  $\boldsymbol{\tau}$ , loadings  $\boldsymbol{\Lambda}$ , as well as the covariances of the latent variables  $\boldsymbol{\Phi}$  and the specific covariances  $\boldsymbol{\Psi}$  and usually take the form of setting individual parameters to a constant (in-

cluding zero - effectively removing the parameter) or constraining a set of parameters to equality. The most basic constraints are most often concerned with the loadings. Instead of just choosing the number of latent variables and assuming that all of them load on all indicators, as was the case in EFA, it is common in CFA to restrict which latent variables load on which indicators and which do not. This is equivalent to constraining the corresponding parameters in  $\Lambda$  to zero while the remaining parameters are freely estimated. As a result, we obtain a loading matrix that is more sparse but obviously less flexible than without the constraints. If the choices with regard to the pre-specification reflect the structure of the true DGP, however, this loss in flexibility will not decrease the goodness-of-fit of the model. This gives way to the core idea of CFA model testing, which will be discussed below.

To further illustrate the implications of the pre-specifications and the constraints they imply, we continue with the example of measuring political ideology. Suppose the group of researchers is not just interested in the most basic distinction between economic left and right positions, but also in a second dimension, often referred to as a cultural dimension. The content and meaning of this second dimension is a topic of debate, but the lowest common denominator is a focus on political issues beyond the economic realm.<sup>2</sup> For simplicity's sake, suppose that the researchers have defined these dimensions appropriately and have created a suitable battery of three survey questions for each construct. For example, using Hooghe et al.'s (2002) definition of a cultural dimension (see footnote 2), questions could include whether respondents believe that climate change is the most urgent question facing our society or whether they're proud to be a member of their country. In slightly more technical terms, the researchers assume  $k = 2$  latent variables for their pre-specification: the left-right dimension ( $\eta_1$ ) and the cultural dimension ( $\eta_2$ ), and  $p = 6$  manifest variables. A reasonable loading structure of the model would therefore be that the first three manifest variables are indicators of  $\eta_1$  and the remaining three are indicators of  $\eta_2$ . Furthermore, it seems plausible that the two latent variables are correlated: In the European context, people with a right-wing position often also hold culturally conservative positions and their support of the left tends to be indicative with more liberal and green positions. However, the proposed structure for the factor loadings indicate that the researchers believe that their manifest variables have been constructed in a way that makes this a pure correlation of the latent variables. In other words, the manifest variables are isolated indicators of these latent variables. Note that this is not a requirement of the model, but a model choice of the researchers. It is possible and there may be good reasons to include manifest variables that are indicators for more than one latent variable. Further, the researchers assume that all interdependence of the manifest variables can be explained by their underlying latent variables. In other words, they are pairwise conditionally independent given the latent variable. Denoting these considerations as model  $\mathcal{M}$ , we can

---

<sup>2</sup>To name but a few conceptualizations, Inglehart (1990) has identified this dimension as one of postmaterialist values, while Kitschelt (1994) defines it as a dimension ranging from libertarian to authoritarian views, and Hooghe et al. (2002) distinguish green/alternative/libertarian from traditional/authoritarian/nationalistic positions.



**Figure 1:** CFA model with two correlated latent variables and six manifest variables.

formally reflect them in  $\Lambda_{\mathcal{M}}$ ,  $\Phi_{\mathcal{M}}$ , and  $\Psi_{\mathcal{M}}$  as follows:

$$\Lambda_{\mathcal{M}} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ & \lambda_{42} \\ & \lambda_{52} \\ & \lambda_{62} \end{bmatrix}, \quad \Phi_{\mathcal{M}} = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \Psi_{\mathcal{M}} = \begin{bmatrix} \psi_{11} & & \\ & \ddots & \\ & & \psi_{66} \end{bmatrix}.$$

To emphasize,  $\Lambda_{\mathcal{M}}$  has this structure because of the assumption that the latent variable  $\eta_1$  loads on the manifest variables 1, 2, and 3 while  $\eta_2$  loads on the remaining manifest variables. The three unique parameters in  $\Phi_{\mathcal{M}}$  are the result of allowing the latent variables to be correlated. Finally,  $\Psi_{\mathcal{M}}$  is a diagonal matrix because of the assumption that all dependencies of the manifest variables can be explained by their respective relationships with the latent variables. This goes to show that the structure of the model requires justification that can only be derived from expert knowledge of the topic at hand.

This structure can, perhaps more intuitively, also be represented graphically. Figure 1 visualizes the model described in the example in accordance with several informal conventions regarding the shapes for latent and manifest variables. Specifically, latent variables are represented with circles, manifest variables with squares and intercepts with a triangle. Furthermore, single-headed arrows represent a linear relationship with a given path coefficient, while the double-headed arrow between  $\eta_1$  and  $\eta_2$  represents their correlation, and lack of an arrow between nodes can be taken to mean (conditional) independence.

### 2.2.2 Estimation

Similarly to the EFA model, parameter estimation of the CFA model is done with a ML approach. Given that the difference between CFA and EFA can be reduced to parameter constraints, all that changes is that these constraints are taken into account in the maximization of the likelihood.

### 2.2.3 Identifiability

Another subtle difference between the EFA and CFA model is how they are rendered identifiable. We have seen earlier that identifiability of the EFA model was achieved with an unrotated solution that confined  $\Phi$  to the identity matrix. For the CFA model, the question of identifiability depends on the set of constraints that are pre-specified by researchers. From a practical point of view, the zero-constraints on  $\Lambda$  will ensure in most cases that there is no issue with rotational invariance. What remains is the issue of scaling of the latent variables for which several alternatives exist (see Little et al., 2006, for an overview). Recall that the location and scale of latent variables is unknown and unobservable. One solution, the marker method comprises of selecting for each latent variable one manifest variable (*marker variable*), for which the loading is set to 1. The result of this approach is, that each latent variable takes the scale of its corresponding marker variable. Note, that this method also solves the issue of rotational invariance from an estimation point of view in case the pre-specified constraints don't. Another approach is *effect coding* (Little et al., 2006), also called *variance standardization*, which for each latent variable, constrains the loadings of all indicators that have a non-zero loading to average 1. In this case, the resulting scale of the latent variables reflects an average of the scales of indicators that is weighted by the magnitude of their respective loadings (Little et al., 2006).

After the scale of the latent variables has been set with a suitable approach, identification is merely a question of model degrees of freedom. Again, these are determined by the number of indicators and the specific constraints imposed on the model. A given CFA model can only be estimated if the number of *free parameters*, i.e. parameters that have to be estimated, doesn't exceed the number of unique pieces of information, also referred to as *knowns*. A model with more knowns than free parameters is called *over-identified*, a model with less knowns than free parameters *under-identified*, and a model with an equal number of knowns and free parameters *just identified*.

Suppose we have  $p$  indicators and a pre-specification for a measurement model  $\mathcal{M}$ . The number of unique pieces of information  $d_{\text{known}}$  is given by the number of sample means and the number of distinct entries in the variance-covariance matrix of indicators. Thus,

$$d_{\text{known}} = \frac{p(p+1)}{2} + p = \frac{p(p+3)}{2}, \quad (15)$$

which is independent of  $\mathcal{M}$ .

On the other hand, the number of free parameters  $d_{\text{free}}$  is given by the sum of the number of intercepts, non-constant factor loadings, non-constant factor covariances, and unique

variances in  $\mathcal{M}$ . Let  $d_{\text{constrained}}$  denote the number of fixed parameters,<sup>3</sup> then

$$d_{\text{free}} = 2p + pk + \frac{k(k+1)}{2} - d_{\text{constrained}}. \quad (16)$$

The degrees of freedom  $d_{\mathcal{M}}$  of our model are then simply given by

$$d_{\mathcal{M}} = d_{\text{known}} - d_{\text{free}}. \quad (17)$$

#### 2.2.4 Testing

Since CFA models are typically estimated via ML, a straightforward test of the model goodness-of-fit can be conducted with a likelihood ratio (LR) test, which can be used for two purposes. First, we can assess the global hypothesis of whether a given model fits the data well. Second, the LR test can be used to compare the fit of two or more (nested) models.

In the first case, we assess the model fit by realizing that a well fitting model should imply a covariance matrix that resembles the sample covariance matrix of the manifest variables. Formally, let  $\mathcal{M}$  denote a model that entails a specification of the structure of all components of the covariance as in the example above. Further, let  $\theta_{\mathcal{M}} := (\mathbf{\Lambda}_{\mathcal{M}}, \mathbf{\Phi}_{\mathcal{M}}, \mathbf{\Psi}_{\mathcal{M}})$  denote the parameters of said model and  $\hat{\theta}_{\mathcal{M}}$  their maximum likelihood estimates (MLE). We can then test the global hypothesis whether the model-implied covariance matrix is equal to the sample covariance matrix

$$H_0 : \Sigma(\theta_{\mathcal{M}}) = \mathbf{S}. \quad (18)$$

Jöreskog (1969) gives a statistic for testing this hypothesis in terms of the fitting function  $F(\cdot)$  as<sup>4</sup>

$$nF\left(\Sigma(\hat{\theta}_{\mathcal{M}}) \mid \mathbf{S}\right) \stackrel{H_0}{\sim} \chi_{d_{\mathcal{M}}}^2. \quad (19)$$

In the second case, where we want to compare two nested models, we can use a standard LR test. Suppose we have a model  $\mathcal{M}_2$  which is nested in model  $\mathcal{M}_1$  and we would like to test the hypothesis

$$H_0 : \Sigma(\theta_{\mathcal{M}_1}) = \Sigma(\theta_{\mathcal{M}_2}). \quad (20)$$

---

<sup>3</sup>Note that constraints arise both from the pre-specification of the model and from constraints placed on the latent variables for identification purposes.

<sup>4</sup>Note, that some sources use a scaling of  $n - 1$  in the statistic. However, both the original paper by Jöreskog (1969) and the prominent implementation of CFA in the R package `lavaan` use the statistic given in (19).

We can test this hypothesis with the LR statistic  $\Gamma$  which is defined as

$$\Gamma(\mathcal{M}_1, \mathcal{M}_2) := 2 \left( F \left( \Sigma \left( \hat{\theta}_{\mathcal{M}_2} \right) \mid \mathcal{S} \right) - F \left( \Sigma \left( \hat{\theta}_{\mathcal{M}_1} \right) \mid \mathcal{S} \right) \right) \stackrel{H_0}{\sim} \chi^2_{(d_{\mathcal{M}_1} - d_{\mathcal{M}_2})} \quad (21)$$

where the distribution holds asymptotically (Wilks, 1938).

[ $\chi^2$ , CFI, RMSEA, find lit review. Start with Rigdon1996, but he doesn't explain why chisq is bad]

### 2.2.5 Factor Extraction

EM algorithm used for fitting CFA in lavaan. M-Step provides a weight method for regression factor extraction? [OPTIONAL] <https://stats.stackexchange.com/questions/126885/methods-to-compute-factor-scores-and-what-is-the-score-coefficient-matrix-in>

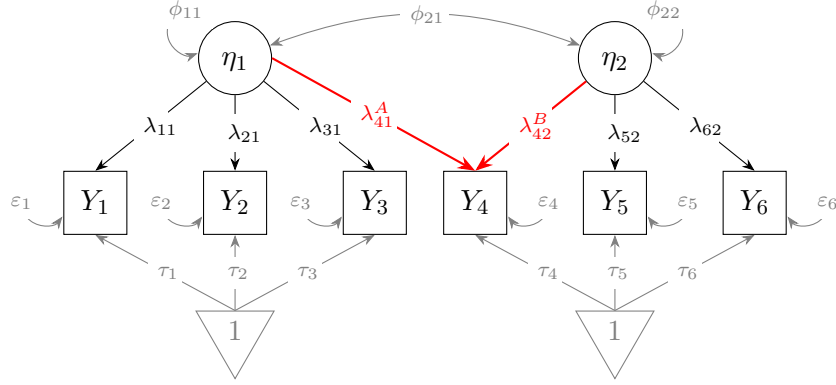
## 3 Measurement Invariance

[TODO: Make clear that configural invariance is not the focus, limit to single latent factor after discussing configural invariance. SHOULD CONFIGURAL INVARIANCE JUST BE A FOOTNOTE?]

To build some intuition of what measurement invariance (MI) is, how violations of it can arise, and what the implications of such measurement non-invariance are for the study of latent constructs in the CFA framework, we return to the example of researchers trying to study political ideology. Then, a formal definition of different types of measurement invariance is given. Finally, the section is concluded with a discussion of how one can test for measurement non-invariance globally and at the item level.

Recall that the group of researchers in the example is interested in studying citizens' positions on a left-right dimension ( $\eta_1$ ) and a cultural dimension ( $\eta_2$ ). To extend the example further, suppose that they are interested in comparing the positions of Citizens across several countries to answer questions such as *do the citizens of Switzerland lean more to the right than German citizens*. It should be easy too see that comparability across groups comes with several caveats that are subsumed in the concept of MI: Fundamentally, such comparisons require the assumption that the pre-specified model structure is equally valid for all (sub)populations under consideration. A manifest variable that is indicative of a latent construct in some populations, but not in others, would constitute an obvious violation of this assumption. For example, it may be the case that a question regarding support for the use of fossil fuels does not relate to the cultural dimension, but to the left-right dimension in a certain country. This difference may be the result of a national public discourse about transitioning out of fossil fuel that revolves more around the loss of jobs than climate change per se. Clearly, the reasons for such structural differences across populations are

manifold and often highly case-specific. Again, we can also visualize this violation. Suppose that for two groups,  $A$  and  $B$ , we have the case described above with regard to  $Y_4$  on the latent constructs. Figure 2 then visualizes how the model structure between the two groups differs. Specifically, the two red arrows emphasize that for group  $A$ ,  $Y_4$  loads on  $\eta_1$ , while for group  $B$ , it loads on  $\eta_2$ . The implications of such a violation is that comparisons of the latent variables across the populations are invalid for the simple reason that they have a different meaning in the respective populations.



**Figure 2:** Violation of configural invariance in the model shown in Figure 1.

Less obvious violations could be that certain indicators are indicative of the left-right position to a different extent across the populations. In this case, the graph would look identical for both populations, but the path coefficients would be population-specific. Suppose that the third item violates MI in this way, so that the true  $\lambda_{31}^A \neq \lambda_{31}^B$ . Ignoring the fact that the groups differ, the researchers obtain something resembling a weighted average of these quantities in their estimation process. Further assume that the populations are absolutely identical in all other respects, including their average position on the latent variables. Given that the group which has a larger true loading on the third item will exhibit a higher average score on that item, the estimated averaged loading across groups will attribute this to the latent construct. Because the same holds in the opposite direction for the other group, the predicted latent positions of the two groups will contrary to the truth exhibit a difference. To relate this more to the concrete example, suppose that an item measures support for increasing or introducing a minimum wage. In Germany, this is a highly politicized issue, while Switzerland doesn't have a federal minimum wage, but instead relies on collective bargaining through unions. Suppose that due to this politicization in Germany, the issue has become strongly related with citizen's left-right position, while in Switzerland, this relationship is much more loose. In other words, the loading in the true model for this item is much higher for Germans. When ignoring this fact in the estimation, the difference between left-right averages across these two populations will be biased. The direction and magnitude of this bias depends on several aspects such as the difference between the group-specific loadings, group sizes, as well as the remaining items.



### 3.1 Types of Measurement Invariance

While these examples have illustrated violations of MI and the implications thereof, the concept of MI may have remained vague. The literature has defined MI in several ways, but “the common denominator of these definitions is a reference to the comparability of measured attributes across different populations” (Davidov et al., 2014, p.58). Put differently, at the core of MI is the question whether a given measurement model measures the same latent variable in a consistent manner across groups. Davidov et al. (2014) stress that this obviously doesn’t imply that there are no differences between the groups on the latent variable. Instead, individuals from different groups with the same position on a latent construct should also be similar with regard to the manifest variables (Davidov et al., 2014; c.f. Mellenbergh, 1989). The importance of MI for CFA therefore stems from the fact that its violation inhibits comparison of the latent variables, which is often the motivation for conducting CFA. Failure to acknowledge this requirement may lead to erroneous results and conclusions about the data. Yet, according to Davidov et al. (2014) tests of MI remain rare in practice despite the fact the violations are common in the cross-national research. Furthermore, the groups across which MI may or may not be violated aren’t necessarily as obvious as in a cross-national context: Violations may also occur at a sub-population level. For instance, measurement non-invariance may arise from differences due to age groups, gender, etc. Clearly, the question of how measurement non-invariance arises in practice is clearly a highly topic-specific question, but in cross-national survey research, an example for an obvious risk for non-invariance would be the translation of survey questions (Davidov & De Beuckelaer, 2010, c.f.). A more general source of measurement non-invariance is the response style of respondents (e.g. Cheung & Rensvold, 2000).

The literature further distinguishes different types of MI, most prominently and importantly configural, metric, and scalar invariance. Following Steenkamp and Baumgartner (1998; c.f. Davidov et al., 2014; Meredith, 1993), these can be seen as hierarchical levels of MI: The most fundamental type, configural invariance, is a prerequisite for metric invariance which is in turn a prerequisite for scalar invariance. In other words, metric invariance is the weaker type of MI and scalar invariance the strongest type which also enables inference on the latent variables to the fullest extent. It is for this reason, that the literature occasionally refers to metric and scalar invariance as *weak* and *strong* invariance, respectively (e.g. Meredith, 1993).

To define these types formally, we first introduce the notion of the multi-group confirmatory factor analysis (MGCFA), which is an extension of the standard CFA model in equation (13). Fundamentally, it is a model of simultaneous and group-specific CFA models for all groups  $l = 1, \dots, g$  such that

$$\mathbf{Y}^l = \boldsymbol{\tau}^l + \Lambda^l \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l. \quad (22)$$

Continuing in this notation, *configural invariance* assumes that the same loading structure is

appropriate across all groups. For configural invariance to hold, the same latent variables must load on the same indicators across all groups, while disregarding differences in magnitude of these loadings. In other words, we require the zero-constrained parameters in the pre-specified structure  $\Lambda$  of the model to hold across groups. Formally, let  $\lambda_{ij}^l$  denote the  $i^{\text{th}}$  manifest variable's loading on the  $j^{\text{th}}$  latent variable in group  $l = 1, \dots, g$ . Then, given the proposed structure  $\Lambda$  across all groups, configural invariance is satisfied if

$$\mathbb{1}_{\{\lambda_{ij}^1=0\}} = \mathbb{1}_{\{\lambda_{ij}^2=0\}} = \dots = \mathbb{1}_{\{\lambda_{ij}^g=0\}} \quad \forall i = 1, \dots, p \ \& \ j = 1, \dots, k, \quad (23)$$

where  $\mathbb{1}_{\{\cdot\}}$  is an indicator function. The attentive reader will have recognized that the first violation of MI in the example above as shown in Figure 2 was a violation of configural invariance. More concretely, configural invariance is violated in the example because

$$\mathbb{1}_{\{\lambda_{41}^A=0\}} = 0 \neq 1 = \mathbb{1}_{\{\lambda_{41}^B=0\}}, \quad (24)$$

$$\mathbb{1}_{\{\lambda_{42}^A=0\}} = 1 \neq 0 = \mathbb{1}_{\{\lambda_{42}^B=0\}}. \quad (25)$$

It should be fairly obvious that the consequences of assuming a common structure  $\Lambda$  across groups when configural invariance is not satisfied may lead to dubious results. There is no guarantee that any detected differences in the latent variables are indeed the result of true differences of the groups. Continuing to take differences at face value ignores the fact that they were obtained from a model which is effectively biased for some or all groups, rendering these differences meaningless. On the flip side, a model which satisfies configural invariance implies that the latent constructs themselves have a comparable meaning across groups as well as the absence of construct bias (Davidov et al., 2014). However, note that this doesn't imply that they are comparable in the quantitative sense of the word.

The next higher level of MI, *metric invariance* can be considered once configural invariance is satisfied. For metric invariance to hold, the factor loadings must be the same across all groups, i.e.

$$\lambda_{ij}^1 = \lambda_{ij}^2 = \dots = \lambda_{ij}^g \quad \forall i = 1, \dots, p \ \& \ j = 1, \dots, k. \quad (26)$$

A model satisfying configural and metric invariance ensures the comparability of the scale of latent variables (Davidov et al., 2014). In other words, metric invariance gives the latent variables a common scale across groups. As a result, the relationships between factor scores obtained from the model and other variables can be compared meaningfully (Davidov et al., 2014; Steenkamp & Baumgartner, 1998). Yet, cross-group comparisons of estimated latent means are still not possible because they may still arise from group-specific item intercepts.

Thus, for the highest level of MI, *scalar invariance*, the intercepts in the CFA model are

required to remain constant across groups, such that

$$\tau_i^1 = \tau_i^2 = \dots = \tau_i^g \quad \forall i = 1, \dots, p. \quad (27)$$

Only if the measurement model satisfies configural, metric, and scalar invariance is it valid to compare latent means across the groups. Given that the errors in the CFA model are assumed to have mean zero, it should be easy to see that only mean differences in the latent factors can result in mean differences of the manifest variables (c.f. Davidov et al., 2014) when all types of MI hold.

For completeness' sake, note that additional types of MI exist. Recall that the item-specific errors  $\varepsilon$  follow a  $p$ -dimensional distribution with  $\text{Cov}(\varepsilon) = \Psi$ . Another type of MI, namely *residual invariance* would then require that the covariances hold across groups, s.t.

$$\Psi^1 = \Psi^2 = \dots = \Psi^g. \quad (28)$$

However, it is obvious that a violation of residual invariance doesn't hinder interpretation of latent means and relationships (Meredith, 1993). Therefore, this thesis only considers configural, metric, and scalar invariance with particular focus on the latter two.

### 3.2 Partial Measurement Invariance

Thus far, MI has been considered as a model property. However, MI can also be viewed as a property of individual items and their corresponding parameters in the CFA model. *Partial measurement invariance* (c.f. Byrne et al., 1989; Steenkamp & Baumgartner, 1998) can thus be seen as the case where the relevant across-group parameter equalities hold for some items, but not for others. As the examples at the beginning of this section have illustrated, this is a natural way of thinking about MI. Cases where all items are either invariant or non-invariant are certainly rather extreme cases. Obviously, MI at the item level and MI at the model level are closely related: The presence of non-invariant items implies non-invariance at the model-level and vice versa.

Byrne et al. (1989) and Steenkamp and Baumgartner (1998) consider partial MI to be achieved when two or more items per latent variable are invariant. Their recommended course of action is then to lift the equality constraints for the non-invariant items, while comparability is ensured by the remaining invariant items. However, the question of whether this is a valid approach for dealing with partial MI remains understudied (Davidov et al., 2014). Nonetheless, full measurement invariance is rarely achieved in practice, even using highly reputable surveys such as the European Social Survey, World Value Survey, or Eurobarometer in cross-national survey research (e.g. Ariely & Davidov, 2011; Davidov, 2008; Ippel et al., 2014). How to proceed under partial MI is therefore a highly relevant question for applied researchers. Instead of the approach above, Davidov et al. (2014) summarize three

options for dealing with partial non-invariance:

1. Restrict analysis to subset(s) of groups for which MI holds
2. Evaluate magnitude of non-invariance and consider removing/replacing items violating MI
3. Study potential sources of non-invariance.

Additionally, scholars have devised methods for eliminating bias which arises from non-invariant items (e.g. Scholderer et al., 2005). Yet, in order to take any of these steps, researchers require information about which groups and items are non-invariant. For group detection, scholars have devised clustering techniques for identifying such subsets of groups (e.g. De Roover et al., 2020; Roover, 2021; Welkenhuysen-Gybels et al., 2007). For the remaining options under partial MI, researchers require additional information about which items are non-invariant. Since this is generally not known in practice, being able to reliably identify these items empirically becomes paramount for enabling valid latent variable comparisons. [Main focus of this thesis]

Although not the focus of this thesis, several comments can be made about these options and how partial MI should be treated more generally. First, the choice of option is highly case-specific and no general recommendation can be made. For example, restricting the analysis to a subset of groups for which MI holds may work perfectly in some cases, but render the analysis irrelevant in others because it may exclude those groups that we want to compare on the grounds of theoretical considerations. Second, these options should not be viewed as mutually exclusive. To the contrary, one should always study or at least consider potential sources of non-invariance. Moreover, the options can and sometimes have to be combined, for instance by subsetting to groups that exhibit less MI and in a next step removing/replacing a non-invariant indicator. Third, the removal of non-invariant items of course restrains the interpretations of the latent variable to which the item referred. As mentioned previously, the selection of items is a crucial design step when devising a CFA model for measuring a latent variable of interest. While the removal of specific items is a rather crude step, it is still an improvement compared to dubious comparisons of latent means from non-invariant models. Ideally, however, scholars are able to replace non-invariant items with comparable, yet invariant, items.

### **3.3 Global Test of Measurement Invariance**

It is possible to test for these types of MI by fitting specific MGCFA models and comparing their goodness-of-fit (Jöreskog, 1971). To build some intuition, we consider two types of MGCFA models: one fully constrained and another fully unconstrained. For the first model, fully constrained, means that each parameter of the CFA model is constrained to equality across all groups. Note that these constraints effectively turn the MGCFA model

into a standard CFA model where groups are altogether ignored. For the second model, fully unconstrained means that each parameter of the CFA model is estimated independently for each group. Note that this flexibility is equivalent to fitting the standard CFA model independently for each group.

The fundamental idea of testing for MI comes from the realization that the fully constrained and the fully unconstrained MGCFA models would be identical under perfect MI and they would exhibit identical goodness-of-fit. Refining this idea, we can sequentially and systematically impose across-group equality constraints for the loadings and intercepts to test whether the goodness-of-fit of the model is significantly worse compared to the multi-group baseline model. If the model fit of these models is similar, we can infer *ceteris paribus* invariance of the constrained family of parameters. To make this more tangible, MI testing typically compares three models: the fully unconstrained model ( $\mathcal{M}_{\text{base}}$ ) which serves as a baseline, the *weakly-constrained model* ( $\mathcal{M}_{\text{weak}}$ ), and the *strongly-constrained model* ( $\mathcal{M}_{\text{strong}}$ ). The fully unconstrained baseline model is constructed as described above. For the weakly-constrained model, we modify the baseline model by constraining all loadings to equality across groups, such that

$$\mathcal{M}_{\text{weak}} : \mathbf{Y}^l = \boldsymbol{\tau}^l + \Lambda \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l, \quad (29)$$

where the omission of the group superscript  $l$  on  $\Lambda$  indicates the equality constraint. Finally, the strongly-constrained model is similarly constructed by modifying the weakly-constrained model with additional equality constraints on the intercepts across groups, such that both the loadings and intercepts are constrained. We thus have

$$\mathcal{M}_{\text{strong}} : \mathbf{Y}^l = \boldsymbol{\tau} + \Lambda \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l. \quad (30)$$

Letting  $\Sigma(\mathcal{M})$  denote the model-implied covariance of model  $\mathcal{M}$ , these three models can be related to three hypotheses corresponding to configural, metric, and scalar invariance. These hypotheses should be tested sequentially because the higher types of MI presuppose that the lower types hold.

$$H_0^{\text{configural}} : \Sigma(\mathcal{M}_{\text{base}}) = \mathbf{S} \quad (31)$$

$$H_0^{\text{weak}} : \Sigma(\mathcal{M}_{\text{weak}}) = \Sigma(\mathcal{M}_{\text{base}}) \quad (32)$$

$$H_0^{\text{strong}} : \Sigma(\mathcal{M}_{\text{strong}}) = \Sigma(\mathcal{M}_{\text{weak}}) \quad (33)$$

As mentioned in the previous section, tests of CFA models have traditionally been conducted with likelihood-ratio (LR) tests because of the prominence of maximum-likelihood estimation of CFA models. Furthermore, it's important to note the nested nature of the three models, which makes using LR tests very convenient. More specifically, using the

LR statistic  $\Gamma$ , defined in equation (21) for comparing two nested models, we have

$$\Gamma(\mathcal{M}_{\text{base}}, \mathcal{M}_{\text{weak}}) \stackrel{H_0^{\text{weak}}}{\sim} \chi_{(g-1)\|\Lambda\|_0}^2. \quad (34)$$

where,  $\|\cdot\|_0$  is the zero "norm" such that  $(g-1)\|\Lambda\|_0$  is the difference in the number of parameters that have to be estimated for the two models.

Analogously, we have for the comparison of the strongly-constrained with the weakly-constrained model

$$\Gamma(\mathcal{M}_{\text{weak}}, \mathcal{M}_{\text{strong}}) \stackrel{H_0^{\text{strong}}}{\sim} \chi_{(g-1)p}^2. \quad (35)$$

These two statistics can then be used to test the hypotheses for MI, formulated above. However, note that for the corresponding metric and scalar invariance to hold, the LR test should fail to reject the respective null hypothesis. Since this is the opposite case of classical hypothesis testing where it's common to reject null hypotheses, the question of statistical power and therefore reasonably large sample sizes becomes crucial (Kim, 2005, c.f.).

While the LR test for MI is popular, scholars have taken issue with this approach (c.f. Drasgow et al., 2018). Their main argument is that with large sample sizes, rejection of the null is almost inevitable because the pre-specified model structure is likely not exactly the true model (e.g. Brannick, 1995; Kelloway, 1995). Of course, this is the case with all statistical tests of point hypotheses and statistically speaking it is questionable whether this really is an issue. In a way, the issue in the literature is the result of a misunderstanding of hypothesis testing. Instead, the truly open question is what constitutes a practically meaningful, not purely statistically significant, change in model fit. Nonetheless, scholars have proposed several alternatives to the LR statistic. Two prominent alternatives, namely the *root mean square error of approximation* (RMSEA; Steiger and Lind, 1980) and the *comparative fit index* (CFI; Bentler, 1990), are introduced in greater detail in the appendix to this thesis. As a rule of thumb, a CFI greater than 0.95 is considered good fit and differences in CFI between two models of more than 0.01 are considered practically relevant (e.g. Cheung & Rensvold, 2002; De Roover et al., 2014).

## 4 Detecting Non-invariant Items

The previous section has formally introduced the concept of measurement non-invariance and has given some intuition how it can arise in research. Further, it was shown that MI can be viewed as a property of the model, but using the framework of partial MI, also as a property of individual indicators in the CFA model. The latter gives researchers the opportunity to remove the group equality constraints for that item or remove (and replace) the item altogether. Naturally, however, it is generally unknown which items are invariant and which are not. An important hurdle is therefore to identify those items

for which MI doesn't hold. This section gives an overview over several methods that have been proposed for this task within the CFA framework. The section then ends with the introduction of a novel method that is highly intuitive and much faster than existing methods.

Before turning to the detection methods, the scope of this and the following sections needs to be limited. First, note that MI at the item level is almost exclusively concerned with metric and scalar invariance. It therefore makes little sense to think of configural invariance as an item-level property, because it is by definition concerned with the overall structure of the model. Thus, configural invariance will be assumed to hold in the following sections. Notwithstanding, many violations of configural invariance would still be detectable with the following methods because they imply metric or scalar non-invariance. Second, note, that there are methods for similar tasks in the IRT literature where MI is better known as differential item/test functioning (DIF) (for an overview, see Tay et al., 2015). However, while there are similarities between IRT and CFA modelling, most of the methods in this literature are not easily transferred to the CFA framework (but see Stark, Chernyshenko, & Drasgow, 2006), leaving me to disregard these alternatives. Third, another alternative way of thinking about non-invariant items is in a Bayesian setting. Here, an alternative exists, where non-invariance of individual item parameters can be explicitly modelled with the help of additional parameters that allow for group-specific deviations from the item parameters (Muthén & Asparouhov, 2013). Finally, I limit myself to a single factor model, i.e. a simple model with a single latent variable and  $p$  indicators. For most methods, this is not strictly necessary. However, it greatly simplifies their implementation as algorithms.

#### **4.1 Existing Methods for Detecting Non-invariant Items**

Scholars have proposed several methods for detecting non-invariant items in the CFA framework. To the best of my knowledge, no systematic comparison of them has been conducted. This may be the result of the fact that this step of the analysis is often considered as something requiring the experience of applied researchers. Thus, especially early methods were rarely formally introduced, but just explained in passing as a part of applied research. For instance, an early way of detecting non-invariant items simply consisted in fitting a fully unconstrained MGCFA and looking for parameters which exhibit large variation across groups (Cheung and Rensvold, 1999; for an application, see e.g. Van de Vijver and Harsveld, 1994). However, as Cheung and Rensvold (1999) point out, the obvious drawback of this method is that it exclusively relies on the researcher's intuition and experience because it is entirely unclear what constitutes a large difference across the groups. Notwithstanding, more formal and test-based methods exist. Another reason for the lack of a systematic review is that the literature is very scattered across many different fields of application. As a result of these two issues, the following list of existing methods may be incomplete.

All of the approaches can be used to identify violations of metric and scalar invariance si-

multaneously. However, this generally doesn't allow for a distinction between which type of MI is violated. One could modify most of the methods to separately detect the type. At the same time, it is questionable how well these approaches would work. Theoretically, the model-comparison approaches would likely show significant differences between the baseline model and both the loading- and intercept-addressing comparison models of a given item that violates only one type of MI. For the simulation study, this would however significantly increase the computational demand, especially for the model-comparison based methods.<sup>5</sup> Because it is certainly more important to be made aware of a violation than to know the exact nature of it, I limit myself to detecting the presence of a violation. Further research could explore how these methods can be accommodated for this task.

#### 4.1.1 Janssens (J)

First, another early method attempts to identify non-invariant items by considering the statistical significance of items. If the loadings for the same item in a fully unconstrained MGCFA model reach statistical significance for some groups, but not for others, they are considered to be non-invariant (Cheung and Rensvold, 1999; for an application, see e.g. Janssens et al., 1995). Obviously, this method can only detect a subset of violations of MI. More concretely, it seems only useful under the assumption that violations of MI only come in the form that some groups deviate from others by having a loading equal to zero with respect to a given item, which is clearly a very restrictive assumption. To illustrate, the method wouldn't be able to identify a non-invariant item for which a sign flip existed across groups as long as the absolute magnitude of that item's loading in both groups was large enough. Moreover, the procedure may be prone to falsely detecting items that exhibit loadings that are close to the critical value of the sample distribution even if the differences across groups are virtually negligible (Cheung & Rensvold, 1999).

Formally, with a single factor model, consider the two null hypotheses of the intercept or the loading being equal to zero in a fully unconstrained MGCFA model for each item  $j$ . Let  $p_{j1}^l$  and  $p_{j2}^l$  denote the corresponding p-values in group  $l$ , for example using a Wald test.<sup>6</sup> For a given significance level  $\alpha$ , item  $j$  is said to violate scalar invariance if  $p_{j1}^l < \alpha$  for at least one  $l$ , but at the same time  $p_{j1}^{l'} \geq \alpha$  for at least one  $l'$ . The same goes for metric invariance by considering the corresponding p-values. Evaluating violations of either type of MI, we have as the set of identified items of this approach

$$S_J := \left\{ j \mid p_{jk}^l < \alpha \ \& \ p_{jk}^{l'} \geq \alpha, \text{ for some } l, l', \text{ and any } k \in \{1, 2\} \right\}. \quad (36)$$

<sup>5</sup>In a baseline model with a single latent variable, the number of models for comparison would double.

<sup>6</sup>Which is what the `lavaan` package provides for each parameter



#### 4.1.2 Modification Indices (MInd)

A second and prominent approach is based on modification indices (MInd). MInd originated from *specification search* which is concerned with achieving a parsimonious and well-fitting model in a step-wise manner (MacCallum, 1986). Generally speaking, MInd measure the increase in model fit that results from inclusion of an additional (group of) parameter(s) to some baseline model. In the context of non-invariant item detection, MInd refer to the additional parameters used when lifting equality constraints for individual indicators in a fully constrained MGCFA model (Cheung and Rensvold, 1999; for an application, see e.g. Riordan and Vandenberg, 1994). A high modification index on a certain parameter therefore suggests that the equality constraint is too restrictive and the item non-invariant. Conveniently, what constitutes high MInd can be quantified because the MInd approach can be framed in a LR test setting which enables the use of significance tests. The obvious drawback is of course that we have to assume that invariance holds for the parameters which remain constrained (Cheung & Rensvold, 1999).

To formalize, let  $\mathcal{M}$  denote the fully constrained MGCFA model and  $\mathcal{M}_j$  a model which is identical except for the modification of lifting the equality constraints on the loadings and intercepts of item  $j$ . The MInd are thus given by the LR statistic of these models

$$\Gamma(\mathcal{M}, \mathcal{M}_j) \stackrel{H_0}{\sim} \chi^2_{2(g-1)}. \quad (37)$$

Let  $t_\alpha$  be the critical value of the  $\chi^2$ -distribution with  $2(g-1)$  degrees of freedom at level  $\alpha$ , then the set of non-invariant items  $S_{\text{MInd}}$  identified by this method is

$$S_{\text{MInd}} := \left\{ j \mid \Gamma(\mathcal{M}, \mathcal{M}_j) > t_\alpha \right\}, \quad (38)$$

which are all items for which we reject the null hypothesis of no difference in goodness-of-fit.

#### 4.1.3 Cheung & Rensvold (CR)

Third, Cheung and Rensvold (1999) have proposed a more elaborate procedure as an extension of an earlier procedure by Byrne et al. (1989). They argue that procedures for detecting non-invariance must take into account the use of marker items which are set to one for the model to be identified. Procedures failing to do so may lead to inaccurate results. To solve this issue, their procedure systematically goes through all pairs of reference and remaining items for which invariance is tested using a *triangle heuristic*. Formally, this procedure begins by specifying a baseline model  $\mathcal{M}$  as a fully unconstrained MGCFA model. This baseline model is then compared with several models for which the loading and intercept of a single item  $i = 1, \dots, p$  is constrained to equality across groups while the remaining parameters remain free to vary across groups. This is repeated while varying

the reference item  $j = 1, \dots, p, j < i$  which is set to 1 for identification purposes (and thus also constrained to equality across groups) in each group of the MGCFA model. As a result, the procedure involves one baseline model and  $p(p-1)/2$  models that are all nested in the baseline model.

Cheung and Rensvold (1999) propose using a  $\chi^2$ -test of these nested models. Formally, we have previously defined the LR-statistic for the difference in the  $\chi^2$ -statistic of  $\mathcal{M}$  and  $\mathcal{M}_{ij}$  as

$$\Gamma_{ij} = \Gamma(\mathcal{M}, \mathcal{M}_{ij}) \stackrel{H_0}{\sim} \chi^2_{2(g-1)} \quad (39)$$

These tests can then be arranged in a strictly lower triangular matrix  $\Gamma$  of test statistics

$$\Gamma := \begin{bmatrix} & & & \\ \Gamma_{21} & & & \\ \vdots & \ddots & & \\ \Gamma_{p1} & \dots & \Gamma_{p(p-1)} & \end{bmatrix}. \quad (40)$$

According to the so-called *triangle heuristic*, proposed by Cheung and Rensvold (1999), this matrix can be simultaneously permuted with a suitable permutation matrix  $P$  by rows and columns to yield a matrix  $\tilde{T} = PTP^T$  which maximizes the number of consecutive rows counted from the first row that include no statistic that exceeds the critical value of the  $\chi^2$ -distribution with  $2(g-1)$  degrees of freedom, given some significance level  $\alpha$ , 0.05 say. In other words, the permutation should rearrange the items such that significant statistics appear in the lower rows of  $\tilde{T}$ . Cheung and Rensvold (1999) then consider those items which appear below the last row that contains no significant test statistics in  $\tilde{T}$  to be non-invariant. More formally, let  $\pi(j)$  be the position of item  $j$  in the permutation yielding  $\tilde{T}$  and let  $l_\alpha$  denote the number of invariant items, as identified by the procedure, then the estimated set of non-invariant items  $S_{CR}$  of this procedure is given by

$$S_{CR} := \left\{ j \mid \pi(j) \leq l_\alpha \right\}. \quad (41)$$

Note that  $\tilde{T}$  is not necessarily unique and by extension  $S_{CR}$  isn't either. Cheung and Rensvold (1999) seem to suggest that - while having a slightly different meaning - all resulting sets are valid. They argue that the choice must be "made in light of substantive issues and underlying theory" (Cheung & Rensvold, 1999, p.12). In my implementation of their idea, I somewhat arbitrarily use the first permutation that contains the maximal number of zero-rows.

#### 4.1.4 Byrne & Van de Vijver (BV)

The fourth and most recent approach by Byrne and Van de Vijver (2010) provides a fairly intuitive and straightforward way of identifying non-invariant items. Similar to the CR and MInd approaches, it is based on model comparisons of goodness-of-fit. However, one crucial difference is the use of the CFI instead of the LR-test. More specifically, the procedure works by first fitting a fully constrained baseline MGCFA model and determining its CFI. Then, for each pair of items and latent constructs with a non-zero loading in the structure of  $\Lambda$ , another model without that relationship is fitted, keeping everything else equal. In a simple measurement model with a single latent variable, the procedure therefore fits the baseline model  $\mathcal{M}$  and  $p$  models  $\mathcal{M}^{(-j)}$  where item  $j$  has been omitted. The intuition is that if item  $j$  is indeed non-invariant, its deletion from the fully constrained model will increase the goodness of fit as measured by the CFI. With regard to a threshold, Byrne and Van de Vijver (2010) consider an item to be non-invariant if its deletion increases the CFI by 0.01 relative to the baseline model.<sup>7</sup> The estimated set of non-invariant items  $S_{BV}$  of this procedure is therefore given by

$$S_{BV} := \left\{ j \mid \text{CFI} \left( \mathcal{M}^{(-j)} \right) \geq \text{CFI}(\mathcal{M}) + 0.01 \right\}. \quad (42)$$

## 4.2 A Novel Approach to Non-invariant Item Detection

[Advantages: Also works if first indicator is non-invariant without awkward reordering of indicators (Compare performance for subsets when  $y_1$  truly is invariant). Fast.] [Disadvantages: ]

The novel approach for detecting non-invariant items deviates from the existing approaches in several ways. First, instead of building on the idea of comparing a baseline model with several alternatives or relying on the model parameters per se, it makes direct use of the implications of the relationship between latent variables and indicators in the CFA model.

To demonstrate the functioning of this novel approach, suppose we have a single latent variable model with  $p$  indicators for the latent variable  $\eta$ . Having estimated such a model, we can obtain estimates for our latent variables  $\hat{\eta}$ . Because the assumed relationship between  $\eta$  and each indicator  $Y_i$  is linear, we can linearly regress  $Y_i$  on  $\hat{\eta}$  and obtain residuals from each regression, denoted  $\hat{\varepsilon}_i$ . More formally, we have

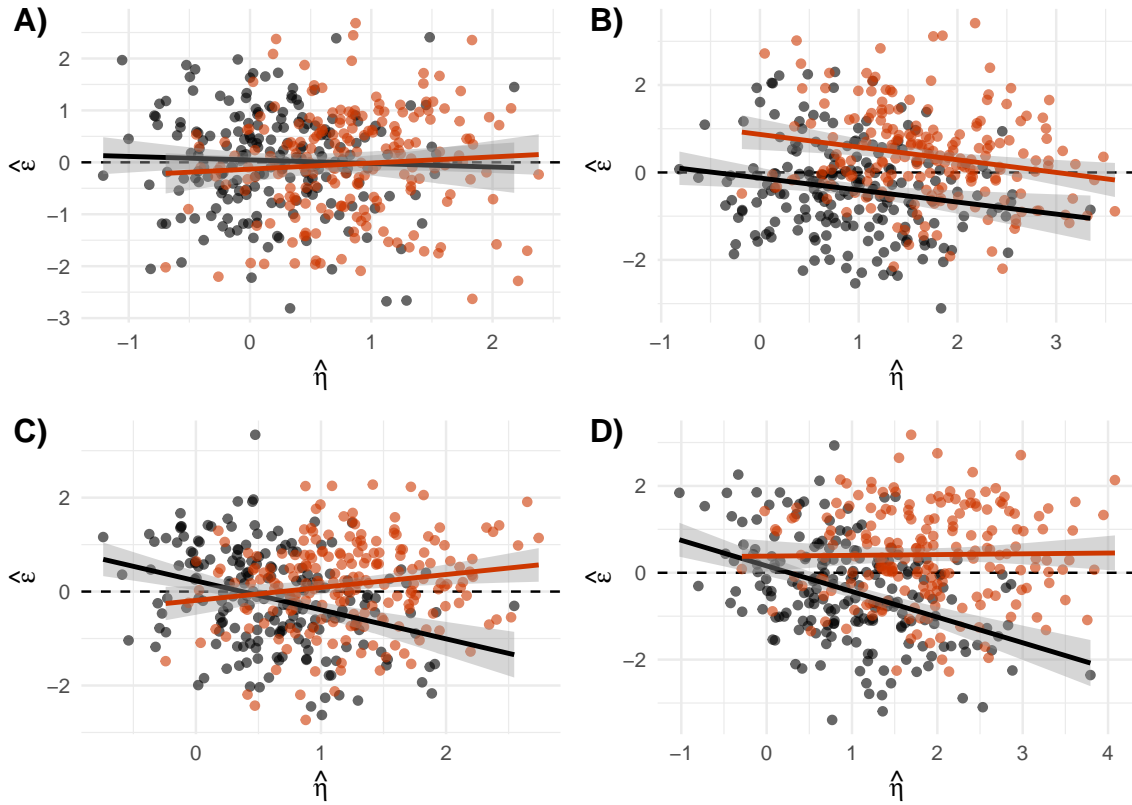
$$\hat{\varepsilon}_i := Y_i - \hat{\mathbb{E}}[Y_i | \hat{\eta}], \quad (43)$$

where  $\hat{\mathbb{E}}[Y_i | \hat{\eta}]$  are the fitted values from regressing  $Y_i$  on  $\hat{\eta}$ . Recall, that, by construction, residuals have mean zero and are uncorrelated with their linear predictor when we're considering all data. Yet, these properties do not necessarily hold for subsets of the full data

---

<sup>7</sup>Justification for the value of 0.01 is provided by Cheung and Rensvold (2002), who consider it to be the critical value for overall measurement invariance.

that was used in the regression. As a result, when analyzing the subset of residuals for a group, the residual mean needn't be zero and the residuals may be correlated with their linear predictor. In fact, if item  $Y_i$  is scalar or metric non-invariant, we would expect a non-zero residual mean or non-vanishing correlation between the residuals and the latent variable for some or all groups. Figure 8 visualizes the differences between residuals from a data-generating process (DGP) that emulates measurement invariance or violations thereof. Panel A shows the setting where invariance is given and the data come from the same DGP and the only difference across groups is a shift in the expectation of the latent variable. In panel B, scalar invariance is violated and the red group's item-specific intercept is shifted up by 0.9 units. As a result, there is a slight difference in the residual mean between the groups and a slight correlation with the latent variable in both groups. In panel C, metric invariance is violated and the red group's loading on  $\eta$  is increased by 0.5 units. Finally, panel D combines both violations. In both panels C and D, there is a substantial correlation between the residuals and the latent variable for the black group.



**Figure 3:** Illustration of the residuals from four DGPs across two groups (black and red).

This example shows that if measurement invariance is violated for a given item, it can be identified by studying the residuals for each group by comparing their means and correlations with the predicted latent variable.

In order to turn this intuition into a procedure, a formalization of mean comparisons and correlations. [DO THAT]

One obvious drawback, that also affects the existing approaches, is that this can only work if  $\hat{\eta}$  is an adequate estimator

The use of residuals from CFA models for diagnostic purposes is by no means an innovative idea. For example Costner and Schoenberg (1973) use correlations of all indicators' residuals to identify relationships that are missing from the specified model.<sup>8</sup> Regardless, to the best of my knowledge, residual analysis has not been used for detecting non-invariance.

### 4.3 Overview over Methods for Detecting Non-invariant items

Method	Type	Reference	Summary
J	Parameter Inspection	Janssens et al. (1995)	
MInd	Model comparison ( $\chi^2$ )	[CITE MIND]	
CR	Model comparison ( $\chi^2$ )	Cheung and Rensvold (1999)	
BV	Model comparison (CFI)	Byrne and Van de Vijver (2010)	
R <sub>1</sub>	Residuals	<i>original</i>	
R <sub>2</sub>	Residuals	<i>original</i>	

Table 2: Caption

## 5 Simulation Study

The simulated data are generated in the same way as those in the partial non-invariance setting in Pokropek et al.'s (2019) simulation study. It generates a single-latent variable model for several groups under varying degrees and types of partial non-invariance. Most of the fixed parameter settings were taken from the original simulation study. I systematically vary seven different parameters, replicating independent simulations for each unique combination of all parameter values summarized in table 3 with the exception on nonsensical combinations such as having more non-invariant indicators than the total number of indicators. In total, this results in XXX simulation settings. In the following, I summarize the relevant aspects of this data generating process.

For each group  $m = 1, \dots, g$ , we begin by generating the true group-specific mean and standard deviation of the latent variable. The  $g$  true latent means  $\mu$  are obtained from a normal

<sup>8</sup>However, they also caution that "this approach can be very misleading" by providing some examples where a modified model is not in line with the true data generating model (Costner & Schoenberg, 1973, p.172)

distribution with mean zero and a standard deviation of 0.3, i.e.  $\mu_m \sim \mathcal{N}(0, 0.3)$ . The  $g$  true standard deviations  $\sigma$  are the absolute value of normal distribution draws with mean one and a standard deviation of 0.1, i.e.  $\sigma_m \sim \mathcal{N}^+(1, 0.1)$ . For simplicity, all groups have equal size  $n$  and these parameters are then used to sample  $n$  positions  $\eta_m$  from a normal distribution for the  $i = 1, \dots, n$  observations nested in  $m$  such that  $\eta_{m(i)} \sim \mathcal{N}(\mu_m, \sigma_m)$ . [Criterion variables?]. In a next step, we sample intercepts  $\tau$  and loadings  $\lambda$  for each of the  $j = 1, \dots, p$  indicators  $Y_j$  from a normal distribution with mean 0 and standard deviation of 0.5 and a uniform distribution on  $[0.65, 0.85]$ , respectively, i.e.

$$\tau_k \sim \mathcal{N}(0, 0.5) \quad (44)$$

$$\lambda_k \sim \text{Unif}(0.65, 0.85). \quad (45)$$

Scores on each indicator  $Y_{m(i)j}$  are finally sampled from a normal distribution with mean  $\tau_j + \lambda_j \eta_{m(i)}$  and standard deviation  $1 - \lambda_j^2$ , i.e.

$$Y_{m(i)j} \sim \mathcal{N}(\tau_j + \lambda_j \eta_{m(i)}, 1 - \lambda_j^2) \quad (46)$$

Note that if it stopped here, the data generating process would reflect perfect MI. To create a setting of partial MI, we first randomly sample  $hg$  groups, where  $h \in \{0.25, 0.5\}$ . Similarly, we randomly sample  $k$  items which will be replaced to exhibit non-invariance. For the affected groups and indicators, the previously sampled intercepts and loadings are altered and used to sample new indicator scores from a normal distribution. The magnitude of this "bias" is set independently for the intercepts and loadings to  $\delta_1$  and  $\delta_2$ , respectively, where  $\delta_1, \delta_2 \in \{0, 0.25, 0.5\}$ . Additionally, the sign of the bias is randomly sampled for each group, indicator, and parameter. Let  $m'$  and  $j'$  denote a group and an indicator that were sampled to be affected by non-invariance. Then

$$Y_{m'(i)j'} \sim \mathcal{N}(\tau_{j'} \pm \delta_1 + (\lambda_{j'} \pm \delta_2) \eta_{m'(i)}, 1 - (\lambda_{j'} \pm \delta_2)^2). \quad (47)$$

Finally, all indicators are discretized to integer values ranging from  $-2$  to  $2$ , using  $\pm 0.47$  and  $\pm 1.3$  as breaks. [justification?]

Parameter		Values	Comment
Number of observations	$n$	{200, 500, 1000}	per group
Number of indicators	$p$	{3, 4, 5, 6}	
Number of groups	$g$	{2, 4, 8, 16}	
Share of affected groups	$h$	{0.25, 0.5}	as share of $g$
Number of non-invariant indicators	$k$	{1, 2, 3}	
Bias on intercepts	$\delta_1$	{0, 0.25, 0.5}	sign of bias randomly sampled for each group and indicator
Bias on loadings	$\delta_2$	{0, 0.25, 0.5}	— " —

**Table 3:** Simulation parameters.

The results of the simulation are in essence classification results for each method. Thus, we can create confusion matrices as well as derive different metrics for evaluating the performance of each method under the different simulation parameter specifications. In the following, I consider a *positive* classification one where an item is identified as non-invariant. Vice versa, a *negative* classification is one where an item is identified as invariant. In combination with the true (non-)invariance of an item, this yields the confusion matrix shown in Table 4 with entries TP (true positive), FN (false negative), FP (false positive), and TN (true negative).

		Predicted	
		non-invariant	invariant
Truth	non-invariant	TP	FN
	invariant	FP	TN

**Table 4:** Confusion matrix.

In the context of this simulation study, the correct identification of truly non-invariant items is paramount. It is arguably much worse when a detection method fails to detect a non-invariant item than if it falsely classifies an invariant item as non-invariant. For example, while falsely removing/replacing invariant items may be theoretically detrimental or costly, it doesn't necessarily invalidate inference on the basis of the latent variables. Therefore, the main performance metric in this case is the *sensitivity* (true positive rate) of the methods, which is defined as the share of correctly identified non-invariant items (TP) among the truly non-invariant items (TP + FN). Nonetheless, a good detection method should also have relatively few falsely identified items. This will be assessed by the secondary performance metric, *specificity* (true negative rate), which is defined as the share of correctly identified invariant items (TN) among all truly invariant items (TN + FP).

- Sensitivity:  $\frac{TP}{TP+FN}$
- Specificity:  $\frac{TN}{TN+FP}$

## 5.1 Results

The remainder of this section summarizes the relevant results and findings from the simulation study. Before going into detail with respect to the effect of different simulation parameters on the performance of the various detection methods, a brief overview across all simulation settings is provided by Table 5. Here,  $N$  denotes the total number of items in all simulation iterations that needed to be classified as either being invariant or non-invariant. With the exception of the BV method, all methods yielded classifications in all parameter settings. The few failures of the BV method all occur in settings with the minimum  $p = 3$  indicators. In these cases, the chance of non-convergence of the likelihood maximization procedure is quite high and thus no CFI can be computed. Note that the omission of these cases somewhat hinders the comparability with the other methods in the otherwise completely randomized design of the simulation study because the omission is not happening at random.

Method	$\delta > 0$			$\delta = 0$	
	N	Sensitivity	Specificity	N	Specificity
J	60,900	0.437	0.628	60,900	0.633
MInd	60,900	0.994	0.319	60,900	0.835
CR	60,900	0.942	0.784	60,900	0.847
BV	59,673	0.785	0.909	60,297	0.995
R1	60,900	0.952	0.641	60,900	0.924
R2	60,900	0.948	0.843	60,900	0.933

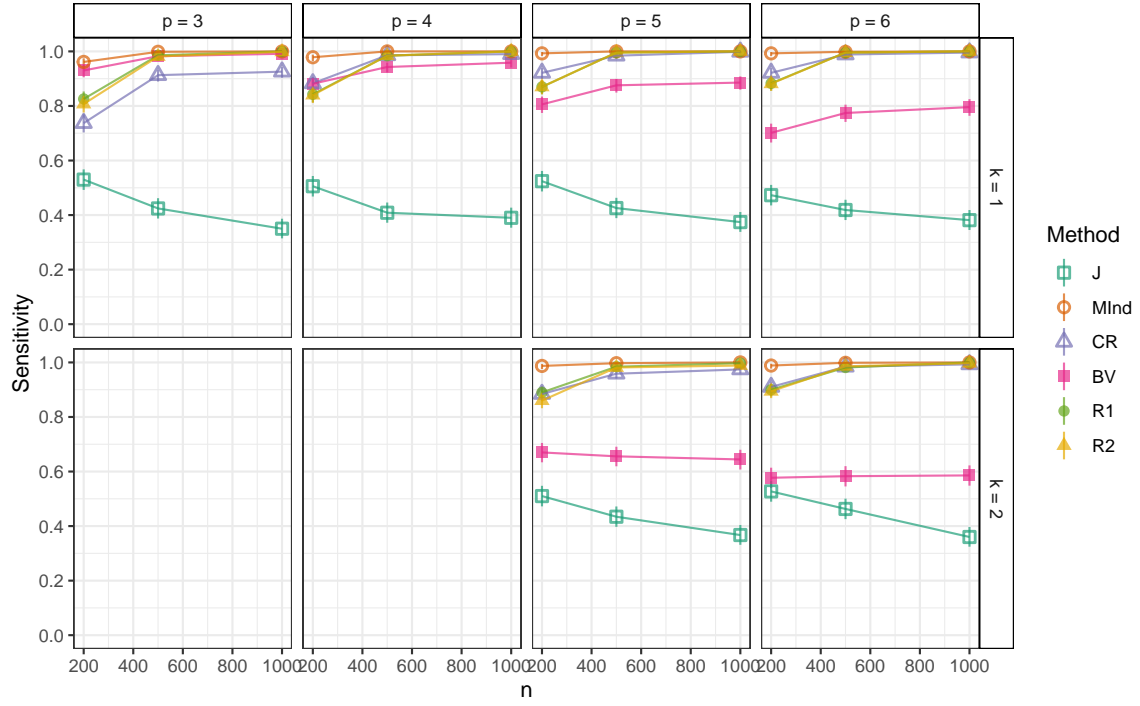
**Table 5:** Caption

With regard to the actual performance of the different method, Table 5 distinguished between the case where there is a substantive bias on the intercepts and loadings. In the former case, where  $\delta > 0$ , all methods but the J method perform adequately, which was to be expected for the reasons discussed in the previous section. With regard to their sensitivity, particularly the MInd, CR, and both R methods performed excellently in detecting well above 90 percent of the items that were truly non-invariant. The highest sensitivity is achieved by the MInd approach, with less than one percent of non-invariant items that were not detected. Yet, this comes at the cost of the lowest specificity of all methods. Likewise, the BV approach exhibits the highest specificity but a relatively low sensitivity. When considering the trade-off between the two metrics, the R2 approach performs best with a very good sensitivity and the second best specificity. When comparing it to the R1 approach, it seems that the step-wise approach yields a substantial increase in specificity for a negligible decrease in sensitivity.

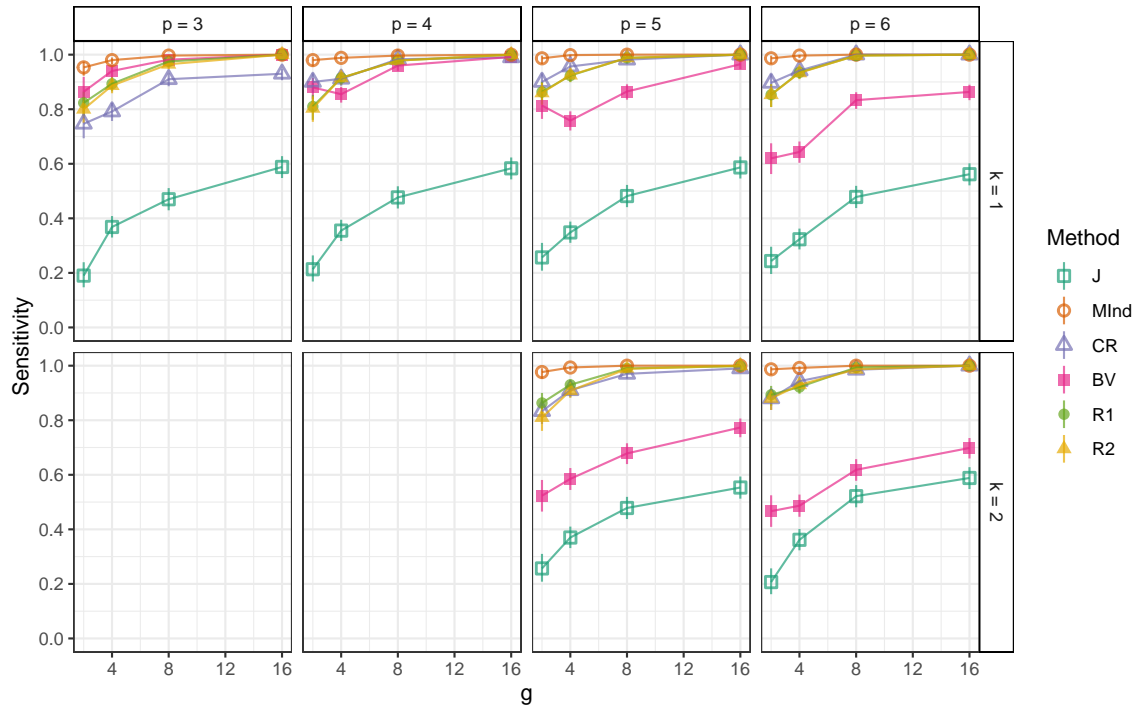
In the perfect MI case, i.e. when  $\delta = 0$ , all methods but the J approach fare decently in classifying the items as negatives.



### 5.1.1 Sensitivity

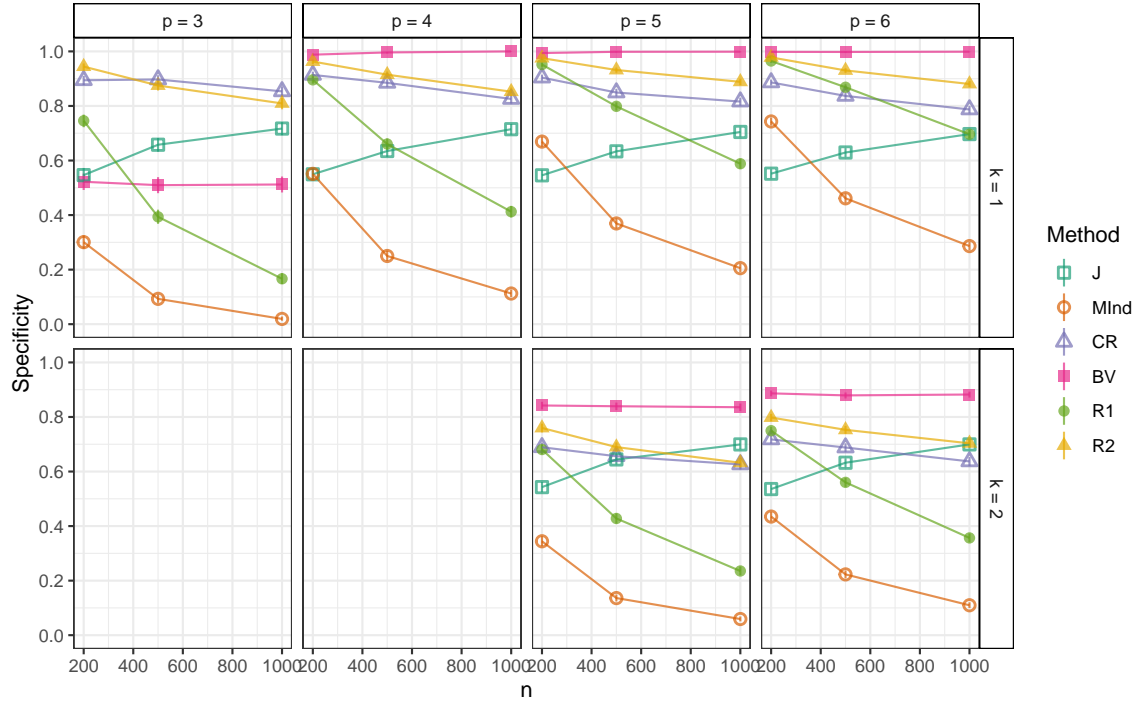


**Figure 4:** Sensitivity of different detection methods as a function of  $n$ ,  $p$ , and  $k$ .

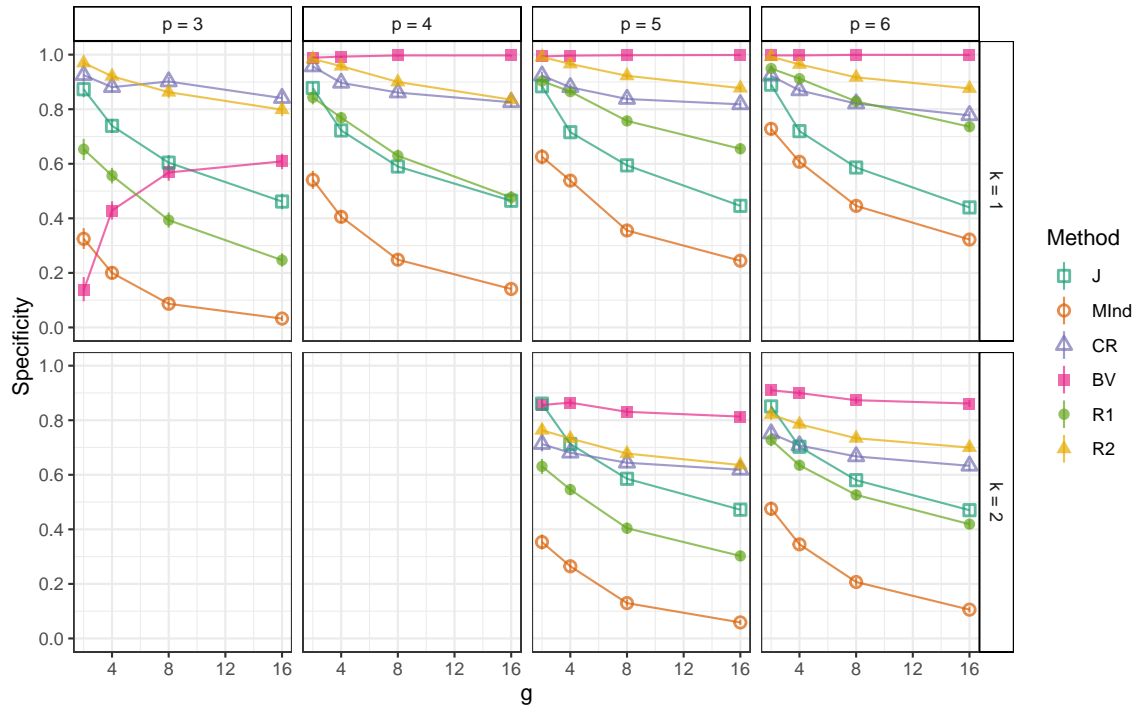


**Figure 5:** Sensitivity of different detection methods as a function of  $g$ ,  $p$ , and  $k$ .

### 5.1.2 Specificity



**Figure 6:** Specificity of different detection methods as a function of  $n$ ,  $p$ , and  $k$ .



**Figure 7:** Specificity of different detection methods as a function of  $g$ ,  $p$ , and  $k$ .

### 5.1.3 Specificity under perfect MI

[NOTE that no sensitivity because TP=0 by construction]

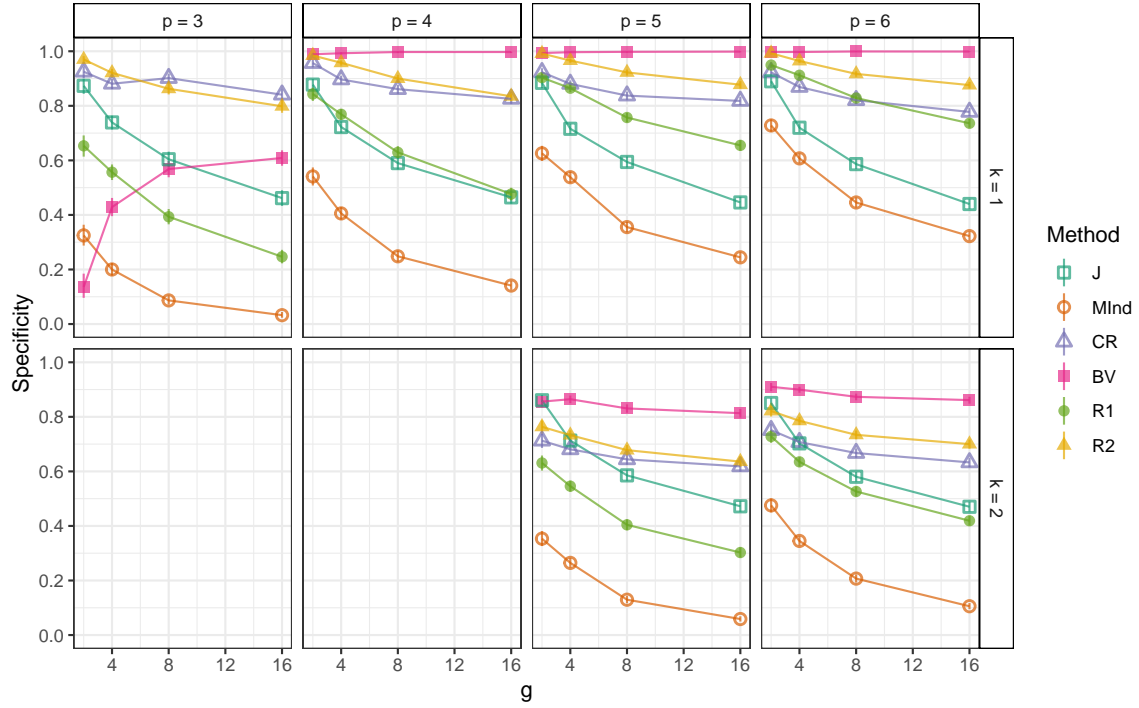


Figure 8: Specificity of different detection methods as a function of  $g$ ,  $p$ , and  $k$ .

## 6 Application: Studying the Cross-National measurement invariance of Populism Scales

### 6.1 Populism Scales

## 7 Conclusion

## References

- Ariely, G., & Davidov, E. (2011). Can we rate public support for democracy in a comparable way? cross-national equivalence of democratic attitudes in the world value survey. *Social Indicators Research*, 104(2), 271–286.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, 107(2), 238.
- Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior*, 16(3), 201–213. <https://doi.org/10.1002/job.4030160303>
- Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research*, 36(1), 111–150.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456.
- Byrne, B. M., & Van de Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10(2), 107–132.
- Cattell, R. B. (1966). The scree test for the number of factors [PMID: 26828106]. *Multivariate Behavioral Research*, 1(2), 245–276. [https://doi.org/10.1207/s15327906mbr0102\\_10](https://doi.org/10.1207/s15327906mbr0102_10)
- Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25(1), 1–27.
- Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *Journal of cross-cultural psychology*, 31(2), 187–212.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233–255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Costner, H. L., & Schoenberg, R. (1973). Diagnosing indicator ills in multiple indicator models. In A. S. Goldberger & O. D. Duncan (Eds.), *Structural equation models in the social sciences* (pp. 168–199). Seminar Press.
- Davidov, E. (2008). A cross-country and cross-time comparison of the human values measurements with the second round of the european social survey. *Survey Research Methods*, 2(1), 33–46.
- Davidov, E., & De Beuckelaer, A. (2010). How harmful are survey translations? a test with schwartz's human values instrument. *International Journal of Public Opinion Research*, 22(4), 485–510.
- Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, 40(1), 55–75. <https://doi.org/10.1146/annurev-soc-071913-043137>
- De Roover, K., Timmerman, M. E., De Leersnyder, J., Mesquita, B., & Ceulemans, E. (2014). What's hampering measurement invariance: Detecting non-invariant items using

- clusterwise simultaneous component analysis. *Frontiers in Psychology*, 5, 604. <https://doi.org/10.3389/fpsyg.2014.00604>
- De Roover, K., Vermunt, J. K., & Ceulemans, E. (2020). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*.
- Drasgow, F., Nye, C. D., Stark, S., & Chernyshenko, O. S. (2018). Differential item and test functioning. *The wiley handbook of psychometric testing* (pp. 885–899). John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9781118489772.ch27>
- Hooghe, L., Marks, G., & Wilson, C. J. (2002). Does left/right structure party positions on european integration? *Comparative Political Studies*, 35(8), 965–989. <https://doi.org/10.1177/001041402236310>
- Horn, J. L. (1967). On subjectivity in factor analysis. *Educational and Psychological Measurement*, 27(4), 811–820.
- Inglehart, R. (1990). *Cultural shift in advanced industrial society*. Princeton University Press.
- Ippel, L., Gelissen, J. P., & Moors, G. B. (2014). Investigating longitudinal and cross cultural measurement invariance of ingelehart's short post-materialism scale. *Social indicators research*, 115(3), 919–932.
- Janssens, M., Brett, J. M., & Smith, F. J. (1995). Confirmatory cross-cultural research: Testing the viability of a corporation-wide safety policy. *Academy of Management Journal*, 38(2), 364–382.
- Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183–202. <https://doi.org/10.1007/BF02289343>
- Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, 36(4), 409–426.
- Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3), 187–200.
- Kelloway, E. K. (1995). Structural equation modelling in perspective. *Journal of Organizational Behavior*, 16(3), 215–224. <https://doi.org/https://doi.org/10.1002/job.4030160304>
- Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 12(3), 368–390. [https://doi.org/10.1207/s15328007sem1203\\_2](https://doi.org/10.1207/s15328007sem1203_2)
- Kitschelt, H. (1994). *The transformation of european social democracy*. Cambridge University Press.
- Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in sem and macs models. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(1), 59–72. [https://doi.org/10.1207/s15328007sem1301\\_3](https://doi.org/10.1207/s15328007sem1301_3)
- MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, 100(1), 107–120.

- Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, 13(2), 127–143.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58(4), 525–543.
- Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis [PMID: 26776378]. *Multivariate Behavioral Research*, 22(3), 267–305. [https://doi.org/10.1207/s15327906mbr2203\\_3](https://doi.org/10.1207/s15327906mbr2203_3)
- Muthén, B., & Asparouhov, T. (2013). Bsem measurement invariance analysis. *Mplus Web Notes*, 17 (Jan 11). <http://www.statmodel.com/examples/webnotes/webnote17.pdf>
- Pokropek, A., Davidov, E., & Schmidt, P. (2019). A monte carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 26(5), 724–744. <https://doi.org/10.1080/10705511.2018.1561293>
- Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective [PMID: 26789208]. *Multivariate Behavioral Research*, 48(1), 28–56. <https://doi.org/10.1080/00273171.2012.710386>
- Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management*, 20(3), 643–671. [https://doi.org/https://doi.org/10.1016/0149-2063\(94\)90007-8](https://doi.org/10.1016/0149-2063(94)90007-8)
- Roover, K. D. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 28(5), 663–683. <https://doi.org/10.1080/10705511.2020.1866577>
- Scholderer, J., Grunert, K. G., & Brunsø, K. (2005). A procedure for eliminating additive bias from cross-cultural survey data. *Journal of Business Research*, 58(1), 72–78.
- Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91(6), 1292.
- Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: Should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, 91(1), 25.
- Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research*, 25(1), 78–90. <https://doi.org/10.1086/209528>
- Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.*

- Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to irt measurement equivalence analysis. *Organizational Research Methods*, 18(1), 3–46.
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.
- Van de Vijver, F. J., & Harsveld, M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the general aptitude test battery. *Journal of Applied Psychology*, 79(6), 852.
- Welkenhuysen-Gybels, J., Van de Vijver, F., & Cambré, B. (2007). A comparison of methods for the evaluation of construct equivalence in a multi-group setting. In G. Loosveldt, B. Swyngedouw, & B. Cambré (Eds.), *Measuring meaningful data in social research* (pp. 357–372). Acco.
- Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62. <https://doi.org/10.1214/aoms/1177732360>
- Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3), 432.

## 8 Appendix

### 8.1 Derivation of the Log-Likelihood of the EFA Model

Recall that the EFA model is

$$\mathbf{Y} = \Lambda \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (48)$$

Assuming that  $\mathbf{Y} \sim \mathcal{N}_p(\Lambda \boldsymbol{\eta}, \Sigma)$ , where  $\Sigma$  is the model-implied covariance matrix

$$\Sigma = \Lambda \Phi \Lambda^\top + \Psi, \quad (49)$$

we can write the likelihood for a single observation  $i$  as

$$\mathcal{L}(\Sigma \mid \mathbf{y}_{1:n}) = (2\pi)^{-p/2} \det(\Sigma)^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{y}_i^\top \Sigma^{-1} \mathbf{y}_i \right\}$$

and the joint likelihood for  $n$  observations as

$$\mathcal{L}(\Sigma \mid \mathbf{y}_{1:n}) = (2\pi)^{-np/2} \det(\Sigma)^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n \mathbf{y}_i^\top \Sigma^{-1} \mathbf{y}_i \right\}.$$

Thus, the joint log-likelihood is given by

$$\begin{aligned} \ell(\Sigma \mid \mathbf{y}_{1:n}) &:= \log \mathcal{L}(\Sigma \mid \mathbf{y}_{1:n}) \\ &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{1}{2} \sum_{i=1}^n \mathbf{y}_i^\top \Sigma^{-1} \mathbf{y}_i. \end{aligned}$$

Next, let us rewrite the last term of the log-likelihood as

$$\begin{aligned} \frac{1}{2} \sum_{i=1}^n \mathbf{y}_i^\top \Sigma^{-1} \mathbf{y}_i &= \frac{1}{2} \sum_{i=1}^n \text{tr} \left( \mathbf{y}_i^\top \Sigma^{-1} \mathbf{y}_i \right) \\ &= \frac{1}{2} \sum_{i=1}^n \text{tr} \left( \mathbf{y}_i \mathbf{y}_i^\top \Sigma^{-1} \right) \\ &= \frac{n}{2} \text{tr} \left( \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top \Sigma^{-1} \right) \\ &= \frac{n}{2} \text{tr} \left( \underbrace{\sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top}_{=: \text{diag}(S)} \Sigma^{-1} \right) \\ &= \frac{n}{2} \text{tr} (S \Sigma^{-1}) \end{aligned}$$



where the first equality holds because the trace of a scalar is equal to the scalar itself, the second holds because of the cyclic property of the trace and the third equality because the sum of traces is the same as the trace of a sum. Finally, note that we can write the diagonal of the sample covariance matrix  $S$  as just  $S$  because it's within the trace function.

Finally, this allows us to rewrite the log-likelihood as

$$\begin{aligned}\ell(\Sigma \mid \mathbf{y}_{1:n}) &= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\Sigma) - \frac{n}{2} \text{tr}(S\Sigma^{-1}) \\ &\propto -\frac{n}{2} (\log \det(\Sigma) + \text{tr}(S\Sigma^{-1}) - \log \det(S) - p)\end{aligned}$$

where the addition of the constants  $\log \det(S)$  and  $p$  conveniently sets the log-likelihood to zero when  $\Sigma = S$ .

### 8.1.1 Connection with the Wishart Distribution

We can show that the likelihood of the factor analysis model above is proportional to that of a  $p$ -dimensional Wishart distribution with the number of observations  $n$  as degrees of freedom and scale matrix  $\Sigma$ . Suppose we have a positive definite matrix  $V \sim \mathcal{W}_p(\Sigma, n)$  with pdf

$$p(v) = \frac{\det(v)^{(n-p-1)/2} \exp\left\{-\frac{1}{2}\text{tr}(v\Sigma^{-1})\right\}}{2^{np/2} \det(\Sigma)^{n/2} \Gamma_p\left(\frac{n}{2}\right)},$$

where  $\Gamma_p(\cdot)$  is the multivariate gamma function. We can thus write the likelihood of our observed sample covariance matrix  $S$  as

$$\begin{aligned}\mathcal{L}(\Sigma \mid S) &= \frac{\det(S)^{(n-p-1)/2} \exp\left\{-\frac{1}{2}\text{tr}(S\Sigma^{-1})\right\}}{2^{np/2} \det(\Sigma)^{n/2} \Gamma_p\left(\frac{n}{2}\right)} \\ &\propto \exp\left\{-\frac{1}{2}\text{tr}(S\Sigma^{-1})\right\} \det(\Sigma)^{-n/2}\end{aligned}$$

which is proportional to the likelihood derived from the joint normal density above.

## 8.2 Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI)

<https://www.tandfonline.com/doi/pdf/10.1080/10705519609540052?needAccess=true>.

## 8.3 Scale invariance of the EFA model

# 9 TEST SECTION - TO BE DELETED

Table A1

	<i>Dependent variable:</i>					
	est					
	J	MInd	CR	BV	R1	R2
	(1)	(2)	(3)	(4)	(5)	(6)
as.factor(n)500	−0.125*** (0.009)	0.028 (0.037)	0.042 (0.033)	0.007 (0.037)	0.043 (0.040)	0.025 (0.041)
as.factor(n)1000	−0.212*** (0.009)	0.052 (0.037)	0.059* (0.033)	0.012 (0.037)	0.066* (0.040)	0.043 (0.041)
as.factor(p)4	0.002 (0.012)	0.009 (0.047)	0.104** (0.042)	−0.068 (0.047)	0.008 (0.051)	0.161*** (0.053)
as.factor(p)5	−0.005 (0.012)	0.010 (0.047)	0.184*** (0.042)	−0.045 (0.047)	0.008 (0.051)	0.282*** (0.053)
as.factor(p)6	−0.003 (0.012)	0.015 (0.047)	0.226*** (0.042)	−0.034 (0.047)	0.010 (0.051)	0.296*** (0.053)
as.factor(k)2	0.003 (0.009)	−0.008 (0.036)	−0.086*** (0.032)	−0.087** (0.036)	−0.010 (0.038)	−0.127*** (0.039)
as.factor(k)3	0.004 (0.010)	−0.011 (0.040)	−0.164*** (0.036)	−0.254*** (0.040)	−0.014 (0.043)	−0.211*** (0.044)
as.factor(h)0.5	−0.021** (0.008)	−0.012 (0.033)	0.025 (0.029)	0.096*** (0.033)	0.001 (0.035)	0.028 (0.036)
as.factor(g)4	0.141*** (0.013)	0.018 (0.052)	0.066 (0.046)	0.030 (0.052)	0.037 (0.056)	0.040 (0.056)
as.factor(g)8	0.286*** (0.013)	0.046 (0.052)	0.131*** (0.046)	0.071 (0.052)	0.072 (0.056)	0.080 (0.056)
as.factor(g)16	0.405*** (0.013)	0.060 (0.052)	0.159*** (0.046)	0.093* (0.052)	0.089 (0.056)	0.105* (0.057)
Constant	0.342*** (0.016)	0.659*** (0.067)	0.412*** (0.060)	0.569*** (0.067)	0.586*** (0.073)	0.390*** (0.073)
Observations	693	693	693	693	693	667
R <sup>2</sup>	0.747	0.007	0.088	0.084	0.009	0.086
Adjusted R <sup>2</sup>	0.743	−0.009	0.073	0.069	−0.007	0.071

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

method	$\delta > 0$			$\delta = 0$	
	n	Sensitivity	Specificity	n	Specificity
BV	59,673	0.785	0.909	60,297	0.995
CR	60,900	0.942	0.784	60,900	0.847
J	60,900	0.437	0.628	60,900	0.633
MInd	60,900	0.994	0.319	60,900	0.835
R1	60,900	0.952	0.641	60,900	0.924
R2	60,900	0.948	0.843	60,900	0.933

**Table A2:** Caption



Eidgenössische Technische Hochschule Zürich  
Swiss Federal Institute of Technology Zurich

## Eigenständigkeitserklärung

Die unterzeichnete Eigenständigkeitserklärung ist Bestandteil jeder während des Studiums verfassten Semester-, Bachelor- und Master-Arbeit oder anderen Abschlussarbeit (auch der jeweils elektronischen Version).

Die Dozentinnen und Dozenten können auch für andere bei ihnen verfasste schriftliche Arbeiten eine Eigenständigkeitserklärung verlangen.

Ich bestätige, die vorliegende Arbeit selbständig und in eigenen Worten verfasst zu haben. Davon ausgenommen sind sprachliche und inhaltliche Korrekturvorschläge durch die Betreuer und Betreuerinnen der Arbeit.

**Titel der Arbeit** (in Druckschrift):

**Verfasst von** (in Druckschrift):

*Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.*

**Name(n):**

**Vorname(n):**


Ich bestätige mit meiner Unterschrift:

- Ich habe keine im Merkblatt „[Zitier-Knigge](#)“ beschriebene Form des Plagiats begangen.
- Ich habe alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert.
- Ich habe keine Daten manipuliert.
- Ich habe alle Personen erwähnt, welche die Arbeit wesentlich unterstützt haben.

Ich nehme zur Kenntnis, dass die Arbeit mit elektronischen Hilfsmitteln auf Plagiate überprüft werden kann.

**Ort, Datum**

**Unterschrift(en)**



*Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich. Durch die Unterschriften bürgen sie gemeinsam für den gesamten Inhalt dieser schriftlichen Arbeit.*