

Department of Mathematics

Master Thesis

Winter 2021/2022

Pit Rieger

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

January 14, 2022

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

Abstract

Violations of measurement invariance (MI) of a given confirmatory factor analysis (CFA) model can arise as a result of non-invariant items and pose a significant threat to the validity of latent variable comparisons across subgroups of a study population. While methods for detecting such items under partial MI exist, there hasn't been a systematic study to compare their performance. This thesis makes three contributions. First, by means of a simulation study, the performance of six detection methods is assessed. Second, two versions of a novel detection approach are introduced and included in the simulation study. The advantage of the novel approach is that it is arguably much easier to interpret than existing methods. Instead of relying on likelihood inference, it builds on residuals and only requires a basic understanding of linear regression, thus being much more accessible to a broad audience of applied researchers. Finally, the detection methods are applied to multiple CFA models for measuring populist attitudes using survey data, demonstrating that they can easily be generalized to fairly complex measurement models. In terms of performance, the findings indicate that one of the existing methods and the novel method can reliably detect non-invariant items. Other existing approaches perform the task they were developed for so poorly that they cannot be recommended under any circumstance. For the exemplary application, the results corroborate findings of significant issues with respect to cross-cultural validity at the model level, but also provide a starting point for model improvement to be taken up by further research.

Further Resources

I provide an interactive version of the simulation study in this thesis as a shiny app hosted at <https://priege.shinyapps.io/miapp/>. The app allows users to specify several parameters for simulating data and estimates the sensitivity and specificity of four detection methods.

Replication files for this thesis can be found on GitHub under <https://github.com/pitriege/masterthesis>.

Contents

1	Introduction	5
1.1	Notation	7
2	Factor Analysis	9
2.1	Refresher: Exploratory Factor Analysis	10
2.2	Confirmatory Factor Analysis	14
3	Measurement Invariance	23
3.1	Types of Measurement Invariance	24
3.2	Partial Measurement Invariance	27
3.3	Global Test of Measurement Invariance	29
4	Detecting Non-invariant Items	32
4.1	Existing Methods for Detecting Non-invariant Items	32
4.2	A Novel Approach to Non-invariant Item Detection	36
4.3	Implementation	41
4.4	Overview over Methods for Detecting Non-invariant items	41
5	Simulation Study	43
5.1	Data Generation	43
5.2	Simulation Setup	45
5.3	Results	46
6	Application: Studying the Cross-National measurement invariance of Populism Models	53
6.1	Synopsis of the Original Paper Castanho Silva et al. (2020)	53
6.2	Implementation of Detection Methods	57
6.3	Results	58
6.4	Summary and Discussion	60
7	Conclusion	62
7.1	Recap	62
7.2	Open questions	62

7.3	Limitations/further research	62
8	Appendix	67
8.1	Derivation of the Log-Likelihood of the EFA Model	67
8.2	Comparative Fit Index (CFI)	69
8.3	Scale invariance of the EFA model	69
8.4	Non-invariance Classifications for Remaining Models in Castanho Silva et al. (2020)	69

1 Introduction

- novel method not able to distinguish between metric and scalar MI. are the others really? reason I didn't implement is because they affect each other.
- write conclusion [NEXT WEEK! next idle task or make notes during reading]
- is weakly, strongly, fully constrained model clear enough? [during next read]
- I vs. we vs. passive [during next read]
- changed order of sections estimation and rotational invariance 02.FA = ζ makes sense [during next read]
- italicize core concepts coherently [extra read]
- check for things that may need reference [extra read]
- check if model structure vs model is sufficiently clear and doesn't clash with "C/EFA model" [extra read] Make sure that \mathcal{M} is the model structure and \mathcal{M}_i or \mathcal{M}' are same model structure with certain additional/removed features e.g. MGCFA or removed item j in BV method

Confirmatory factor analysis (CFA) (Jöreskog, 1969) is widely used to infer latent concepts, i.e. quantities of interest that cannot be observed directly. In survey research, CFA models are used to generate an estimate of a latent variable on the basis of a battery of question items relating to the concept. Prominent examples of latent concepts include mathematical skills in standardized tests, personality traits or happiness in psychology, and ideology or trust in a political system in political science. Oftentimes, researchers are interested in comparisons of latent variables across different groups or sub-populations within the general study population. What constitutes a group depends on the context of the study and can be as diverse as countries, questionnaire language, gender, age groups, or time. Comparisons of latent variables across such groups come with the crucial – yet often overlooked – caveat of measurement invariance (MI, also referred to as measurement equivalence) (for an overview, see Davidov et al., 2014). In essence, MI is the requirement that the items that relate to the latent variables in the CFA model function in the same way in all groups within the study population. In the context of survey research, this implies that the included questions must be taken in and answered in all groups alike. Violations of MI can arise for numerous reasons but they are generally the result of at least one group responding to at least one question in a systematically different way, irrespective of their actual position on a latent variable. The implication of undetected violations is that the estimates of the latent variable are biased because the true underlying model is not identical for all groups. As a result, violations of MI can render differences on the estimated latent variable completely meaningless.

Several tests for MI are available and fall into one of two categories. First, global tests aim to establish whether there is an issue with regard to MI anywhere in the model. Second, tests at the group- or item-level aim to identify which specific groups or items function differently. In other words, the second category is concerned with the more fine-grained question of which items or groups contribute to global non-invariance. This master thesis

deals with approaches to detect items that violate MI, i.e. in the context of survey research questions which are systematically understood or answered differently across groups. These tests are particularly useful in the setting of partial MI (Byrne et al., 1989) under the assumption that MI holds for a subset of groups or items because they point to which modifications are necessary to achieve global MI. With this information, researchers can improve their measurement models in terms of their cross-group validity, for example by replacing or removing items that contribute significantly to measurement non-invariance. I discuss existing methods for the detection of non-invariant items and make an original contribution by proposing two variants of a novel approach. The advantage of the new method stems primarily from the fact that it's built on the resemblance of CFA with linear regression. I argue that residuals from CFA models can be studied to detect violations of MI at the item level. This is the distinguishing characteristic compared to existing approaches which generally rely on likelihood-based goodness-of-fit tests. The gain from instead devising a new method on the basis of residuals is threefold. First, linear regression and residuals are well understood and intuitive. Statistical tests of different properties of the residuals are thus straightforward. This advantage is particularly valuable because CFA is primarily used in applied research which has in part neglected questions of MI (c.f. Davidov et al., 2014). Second, the novel approach can easily be visualized. Again, this is helpful for promoting the use of this detection method among applied researchers. Finally, contrary to goodness-of-fit-based tests, less models need to be fitted which results in the method being less demanding, both cognitively and computationally. For some of the existing approaches, the number of models that need to be fit increases exponentially in the number of items. Instead, for the new approach, it can be as low as one, but in any case lower than for existing methods.

Another contribution of this thesis to the literature on MI is the systematic comparison of both existing and new approaches by means of a simulation study. To the best of my knowledge, this is a gap in the literature. The simulation study in this thesis thus helps answering the open question how well the existing methods actually work and provides a benchmark against which my novel approach can be compared. What's more, I include an application of the existing and novel methods for CFA models used by political scientists in the study of populist attitudes in a cross-country setting. More specifically, I apply the methods to seven models, with particular focus on two, using replication data from Castanho Silva et al. (2020) who asked respondents a broad range of items necessary for these models. This serves as a proof of concept by showing that these methods can in principle be applied to fairly complicated CFA models using real-world data. Additionally, it is of substantive interest to see which models suffer from violations of MI and which survey items are particularly problematic. Since model development is at least in part driven by empirical considerations (Castanho Silva et al., 2020), this application shows how detection methods at the item level can be used as a starting point for model development.

The findings of the simulation study suggest that only one of the four existing detection methods, the BV method, makes sufficiently sensitive and specific item classifications. The

remaining three methods perform rather poorly, often due to very low specificity. At the same time, one version of the novel approach introduced in this thesis is a serious contender to be the best performing detection method. While it is slightly less specific than the BV method, it achieves noticeably higher sensitivity. This seems like a good trade-off because it is arguably worse to overlook a truly non-invariant item than to falsely identify an invariant item as being non-invariant. Furthermore, it is very encouraging to see that both the existing and novel best-performing methods are the most straightforward ones in terms of theory and implementation.

In the application of the models to the populism CFA models, the results are that there can be considerable disagreement the different methods. This isn't particularly surprising given the poor global MI properties of the models as documented by Castanho Silva et al. (2020). Nonetheless, the generalization of detection methods to CFA models with multiple latent variables was straightforward and shows that they are not limited to single-factor models. Furthermore, for the two models analyzed in greater detail, clear sets of items that are likely major contributors to the global MI of the models could be identified. These may serve as the starting point for attempts to improve these existing populism models with particular focus on their MI properties for cross-national comparisons.

This thesis progresses as follows. First, I review the fundamentals of exploratory factor analysis (EFA) in order to then introduce and distinguish the CFA model from it. Next, I define the concepts of MI and partial MI. I discuss how measurement non-invariance, i.e. a violation of MI, can arise in the CFA framework and discuss its implications for the measurement and analysis of latent variables. I then introduce the existing methods for detecting non-invariant items as well as my novel approach. These are put to the test in the subsequent simulation study. Before turning to a final discussion and conclusion, I apply all methods to real-world data of populist attitudes in a cross-national context.

1.1 Notation

The notation in this thesis mostly follows the conventions in the relevant literature on CFA and MI. Vectors are column vectors. Both vectors and matrices are written in boldface. Estimates are denoted with a hat symbol ($\hat{\cdot}$). Nested indexes are written in parentheses, e.g. if index i is nested in l , it is written as $l(i)$. An overview of the unique symbols used in this thesis is given for reference in table 1 below. Rarely, the same symbols are used to refer to different things, e.g. index i is generally used for items $i = 1, \dots, p$, but sometimes for observations $i = 1, \dots, n$. These cases are always clearly documented.

Specific CFA model (structure)	\mathcal{M}
Items / indicators	\mathbf{Y}
Latent variables	$\boldsymbol{\eta}$
Intercepts	$\boldsymbol{\tau}$
Errors / specific variables	$\boldsymbol{\varepsilon}$
Loading matrix	$\boldsymbol{\Lambda}$
Loadings	λ_{ij}
Covariance matrix of errors	$\boldsymbol{\Psi}$
Covariance matrix of latent variables	$\boldsymbol{\Phi}$
Covariance matrix of items	$\boldsymbol{\Sigma}$
Sample covariance matrix of items	\mathbf{S}
Number of items	p
Number of latent variables	k
Number of groups	g
Number of observations per group	n
Total number of observations	N
Item index	$i = 1, \dots, p$
Latent variable index	$j = 1, \dots, k$
Group index	$l = 1, \dots, g$
Degrees of freedom	d
Significance level	α
p-value	\wp
Magnitude of bias on intercepts	δ_1
Magnitude of bias on loadings	δ_2
Fit function	$F(\cdot)$
Likelihood function	$\mathcal{L}(\cdot)$
Log-likelihood function	$\ell(\cdot)$
Likelihood ratio statistic	$\Gamma(\cdot)$
Set of non-invariant items	\mathbf{S}
Regression intercept	γ
Regression coefficient	β
Latent variable means	$\boldsymbol{\mu}$
Latent variable standard deviations	$\boldsymbol{\sigma}$
Indicator function	$\mathbb{1}_{\{\cdot\}}$
Normal distribution	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$
Multivariate-normal distribution	$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Identity matrix	$I_{p \times p}$

Table 1: Overview of symbols used in this thesis

2 Factor Analysis

Many scientific concepts that are of interest cannot be observed directly. This is especially true in the social sciences. Quantitative researchers usually refer to such concepts as *latent* variables. 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 for measurement models. In a nutshell, CFA measurement models circumvent unobservability by utilizing several observable *items* that (are assumed to) relate to the latent variables. Under linearity of these relationships, factor analysis can be interpreted as a model of the covariance of the items by assuming that latent variables account for these covariances.

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 implications of using CFA when inferring latent variables. Before going into the formal introduction of the CFA framework, the well-known *exploratory factor analysis* (EFA), which turns out to be a special case in the CFA framework, is recapitulated. This makes it much easier to then elaborate on the subtle changes when generalizing EFA to yield the CFA framework. For both, 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, an 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 an obvious example of a latent construct because it cannot 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 the 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 assume a structure of how these items relate to the latent variable and to one another with a measurement model in the CFA framework. Ultimately, such models can then be used

to infer respondents' ideological positions.

This motivating example also highlights several non-technical fundamental implications of using CFA. These will not play a large role in the formal introduction below owing to the fact that the specification of measurement models is typically done on the basis of expert knowledge and theoretical considerations, making it highly field-dependent. Notwithstanding, model decisions regarding both the model structure and the choice/construction of items are highly consequential for the interpretation of the inferred latent variables. First, the construction and choice of a set of items 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 items in a way that reflects and covers the entire concept. Second and in a similar spirit, the concrete specification of the measurement model, including various aspects such as the number of latent variables, their relationships both with other latent variables and the items, etc. has a significant impact on the model estimates and in turn interpretation and prediction.

2.1 Refresher: Exploratory Factor Analysis

2.1.1 Setup

EFA supposes we have a set of p continuous¹ items $\mathbf{Y} = (Y_1, \dots, Y_p)^T \in \mathbb{R}^p$, also referred to as *manifest variables* or *indicators*, that are assumed to relate to k latent variables $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)^T \in \mathbb{R}^k$ 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, in the practical application of EFA, 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 $\boldsymbol{\tau} = (\tau_1, \dots, \tau_p)^T$ are intercepts, λ_{ij} are regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)^T$ are errors. In the factor analysis, we refer to the regression coefficients as (*factor*) *loadings* and the errors as *specific variables*. Note, that the assumed relationship implies that each item is a linear combination of all latent variables plus an intercept and an idiosyncratic error.

¹Strictly speaking, items are rarely – if ever – measured on continuous scales. However, particularly in survey research, it is common to treat response scales with five or more categories as continuous (c.f. Pokropek et al., 2019). Moreover, several studies have shown that, using maximum likelihood estimation, this practice yields valid results (e.g. Johnson & Creech, 1983; Muthén & Kaplan, 1985)

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

$$\mathbf{Y} = \boldsymbol{\tau} + \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (2)$$

2.1.2 Assumptions

To emphasize, $\mathbf{\Lambda}$ is unknown and $\boldsymbol{\eta}$ is unobservable. As demonstrated by the motivating example, this is the fundamental reason for conducting factor analysis. In order to obtain estimates for these quantities, we thus 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 error assumptions in linear regression. Furthermore, they allow the decomposition of the covariance matrix of \mathbf{Y} as

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

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

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

where $\boldsymbol{\Phi} := \text{Cov}(\boldsymbol{\eta})$. It is clear from this decomposition that by assuming a relationship of the form shown 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 $\boldsymbol{\Sigma}(\mathbf{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ to emphasize this dependence. This is crucial for the estimation of factor analysis models: A good model should resemble the observed sample covariance matrix of the items. As the next subsection demonstrates, the model parameters can consequently be estimated by minimizing the difference between the sample and model-implied covariance. Similarly, the model-implied covariance of a fitted model can be used to establish the goodness-of-fit of the model by comparing it with the sample covariance matrix.

2.1.3 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

$$\tilde{\Lambda} := \Lambda \mathbf{R} \quad (5a)$$

$$\tilde{\eta} := \mathbf{R}^{-1} \eta \quad (5b)$$

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

$$\mathbf{Y} = \tilde{\Lambda} \tilde{\eta} + \varepsilon = \Lambda \mathbf{R} \mathbf{R}^{-1} \eta + \varepsilon = \Lambda \eta + \varepsilon \quad (6)$$

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 Λ and η . Although \mathbf{R} is not strictly limited to rotation matrices, it is called a rotation and the aforementioned property is referred to as rotational invariance.

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

$$\mathbb{E}[\eta] = 0 \quad (7a)$$

$$\text{Cov}(\eta) = I_{k \times k}, \text{ an identity matrix.} \quad (7b)$$

Given these properties of η , it is easy to see that $\text{Cov}(\tilde{\eta}) = I_{k \times k}$ 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.² 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

²Note that the constraints are arbitrary because we cannot know the true location and scale of the latent variables. Instead, we arbitrarily set them for technical purposes.

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). Transformed or untransformed, factor analysis always requires interpretation of the latent variables. However, the fact that EFA doesn't have a unique solution has provoked criticism for the supposed subjectivity involved in rotations (e.g. Horn, 1967; but also see Mulaik, 1987).

2.1.4 Estimation

EFA models are typically estimated by means of a maximum likelihood (ML) approach³ which requires an additional distributional assumption. The usual assumption of a multivariate normal distribution can be written as

$$\mathbf{Y} \sim \mathcal{N}_p(\boldsymbol{\tau} + \mathbf{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Sigma}). \quad (8)$$

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

$$\tilde{\mathbf{Y}} \sim \mathcal{N}_p(\mathbf{\Lambda}\boldsymbol{\eta}, \boldsymbol{\Sigma}), \quad (9)$$

where in practice, we simply subtract the corresponding sample mean from each item to ensure their centering due to the additional assumption that the latent variables have mean zero.

Estimates can then be obtained by maximizing the normal log-likelihood over the parameters in $\mathbf{\Lambda}$ and $\boldsymbol{\Psi}$. It is easy to overlook how astonishing it is that we can obtain estimates for these quantities given that $\boldsymbol{\eta}$ is unknown. This is due to the assumptions about $\boldsymbol{\eta}$ which render the likelihood dependent on only $\boldsymbol{\Sigma}$ which in turn depends on the quantities of interest $\mathbf{\Lambda}$ and $\boldsymbol{\Psi}$.

Let $\theta := (\mathbf{\Lambda}, \boldsymbol{\Psi})$, then the log-likelihood is given by

$$\ell(\theta \mid \mathbf{S}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log \det(\boldsymbol{\Sigma}(\theta)) - \frac{n}{2} \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\theta)), \quad (10)$$

where \mathbf{S} is the sample covariance matrix of \mathbf{Y} . It turns out that \mathbf{S} is a sufficient statistic for the parameters in the EFA model: 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 minimiz-

³Another 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.

ing the fit function

$$F(\theta | \mathbf{S}) = \log \det(\boldsymbol{\Sigma}(\theta)) - \log \det(\mathbf{S}) + \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\theta)) - p \propto \ell(\boldsymbol{\Sigma}(\theta) | \mathbf{S}), \quad (11)$$

where the replacement of constants in equation (10) with $\log \det(\mathbf{S})$ and p conveniently sets the fit function to zero when $\boldsymbol{\Sigma}(\theta) = \mathbf{S}$, i.e. when the model fits perfectly.

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. In the following, we will see that the role of k is one fundamental difference between EFA and CFA.

2.2 Confirmatory Factor Analysis

2.2.1 Setup

As mentioned previously, the CFA framework can be viewed as a generalization of EFA. At first glance, the fundamental setup thus remains identical: The goal is still to try to model the continuous items as a linear function of latent variables. We can keep all prior notation and still write the model in matrix notation as

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

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

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

The consequential difference lies in how we view the parameters in CFA: To formulate a CFA model, researchers must define the *structure* of their model. By structure, we refer to various aspects in the CFA framework relating to any of its parameters. The structure defines which relationships and covariances are possible when fitting the model by determining which parameters are freely estimated and which are set to zero, effectively removing them, or subjected to other constraints. Decisions which yield a model structure are typically justified theoretically and with substantive knowledge of the subject at hand. For example, we may want to formulate a model structure where items load only on a subset of latent variables. Recall that for EFA we simply allowed for every item to load on all k latent variables. At a minimum, any CFA model structure needs to specify the number of latent variables (k), which items load on which latent variables (Λ), which latent variables are dependent on each other (Φ), and which specific variables are dependent on each other (Ψ). The implications of a given model structure can be viewed as a set of constraints on some of the parameters in the CFA framework while others remain free. In the simplest and perhaps most common case, these constraints are zero constraints that effectively remove relationships or covariances from the model by forcing some loadings or covariances to zero. For example, assuming that a certain item isn't related to one of the latent variables, i.e. doesn't load on said latent variable, is equivalent to constraining its loading to zero.⁴ Similarly, assuming pairwise uncorrelated latent variables is equivalent to constraining all parameters that are not on the diagonal of Φ to zero. It should now be easy to see that EFA can be represented as a special case in the CFA framework. It is a CFA model where we assume the following model structure: All p items load on all k latent variables which are pairwise uncorrelated and the specific variables are also pairwise uncorrelated. For the unrotated solution, we further assumed that Φ is an identity matrix.

Although model structures typically just define possible relationships and covariances, they may be much more specific. Note, that the above minimum components of a model structure are ignorant as to the magnitude of the parameters that are not constrained to zero. However, a model structure may also include the fixing of free parameters to constants. For example, we may constrain a single loading to be 2. However, such constraints are rare in practice because they are difficult to justify.⁵ What is more common are equality constraints where two or more parameters are assumed to be equal. For example, we may want to ensure that two loadings have the same magnitude. One may ask why we would ever want to make such assumptions about a model and impose the corresponding constraints. The general idea is that model structures can be thought of as hypotheses about the data-generating process of the constituting items. In a CFA framework, theoretically generated or justified model structures can then be tested or "confirmed" empirically – hence the name *confirmatory* factor analysis. Model testing in the CFA framework will be

⁴However, note, that this doesn't imply that they are uncorrelated because they may still be connected through some other variable.

⁵An exception is the marker variable which can be used for model identification purposes and is discussed in greater detail below. However, the marker variable is a technical solution to the problem of latent variables having no unique scale. The real issue alluded to here arises when the scale of a latent variable is fixed and we then impose a constant loading on the relationship between said latent variable and another item.

discussed in greater detail below, but the general idea is to compare model-implied covariance matrices with the observed sample covariance matrix. If the choices with regard to the model structure resemble the structure of the true data generating process (DGP), the model will have better goodness-of-fit than otherwise. Another advantage of this flexibility is that it enables researchers to incorporate their substantive field knowledge into their models. Yet, this flexibility cuts both ways: Formulating good CFA models is hard and requires detailed knowledge of the subject matter. 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 a question of substantive theoretical considerations rather than statistics. As such, it varies heavily across disciplines and fields.

To further illustrate the implications of model structures 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. 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. 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, 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 citizen of their country. In slightly more technical terms, the researchers assume $k = 2$ latent variables for the structure of their model: the left-right dimension (η_1) and the cultural dimension (η_2), and $p = 6$ manifest variables. A reasonable loading structure of the model would therefore be that the first three items are indicators of η_1 and the remaining three are indicators of η_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 items are isolated indicators of these latent variables. Note that this is not a requirement of CFA models in general, but a model choice of the researchers. It is possible and there may be good reasons to include items that are indicators for more than one latent variable. Further, the researchers assume that all interdependence of the items 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

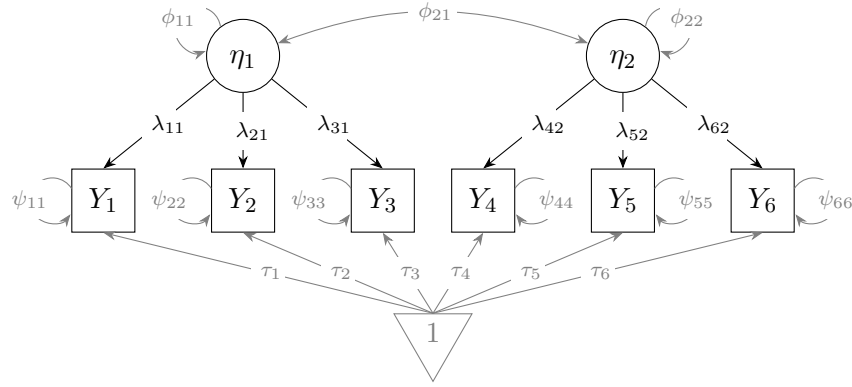


Figure 1: CFA model with two correlated latent variables and six manifest variables.

\mathcal{M} , we can formally reflect them in $\mathbf{\Lambda}_{\mathcal{M}}$, $\mathbf{\Phi}_{\mathcal{M}}$, and $\mathbf{\Psi}_{\mathcal{M}}$ as follows:

$$\mathbf{\Lambda}_{\mathcal{M}} = \begin{bmatrix} \lambda_{11} \\ \lambda_{21} \\ \lambda_{31} \\ & \lambda_{42} \\ & \lambda_{52} \\ & \lambda_{62} \end{bmatrix}, \quad \mathbf{\Phi}_{\mathcal{M}} = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \mathbf{\Psi}_{\mathcal{M}} = \begin{bmatrix} \psi_{11} & & \\ & \ddots & \\ & & \psi_{66} \end{bmatrix}.$$

To emphasize, $\mathbf{\Lambda}_{\mathcal{M}}$ has this structure because of the assumption that items 1, 2, and 3 load on the latent variable η_1 while the remaining items load on η_2 . The three unique parameters in $\mathbf{\Phi}_{\mathcal{M}}$ are the result of allowing the latent variables to be correlated. Finally, $\mathbf{\Psi}_{\mathcal{M}}$ 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.

What's neat about model structures of CFA models is that they can be represented graphically. To illustrate, 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 η_1 and η_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 maxi-

mization of the likelihood: For example, parameters that are constrained to zero are held fixed at zero for the ML estimation. Similarly, we estimate a single parameter for two or more parameters that are constrained to equality by the model structure.

2.2.3 Identifiability

Another subtle difference between the EFA and CFA model is how they are rendered identifiable. Recall that identifiability of the EFA model depends on k and is achieved with the help of an unrotated solution that confines Φ to the identity matrix. For the CFA model, the question of identifiability depends on the set of constraints that researchers pre-specify in their model structure. From a practical point of view, the zero-constraints on Λ 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 as a result of latent variables being unobservable, 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 structural 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 items 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 items 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 which are determined by the structure of 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 items in a model structure which we denote \mathcal{M} . The number of unique pieces of information d_{known} is given by the number of sample means and the number of distinct entries in the variance-covariance matrix of items. 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_{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,⁶ then

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

The degrees of freedom for the given model structure $d_{\mathcal{M}}$ are then

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

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.

For the first case, realize 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 $\theta_{\mathcal{M}} := (\Lambda_{\mathcal{M}}, \Phi_{\mathcal{M}}, \Psi_{\mathcal{M}})$ 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 $F(\cdot)$ as⁷

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

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

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

⁶Note that constraints arise both from the structure of the model and from constraints placed on the latent variables for identification purposes.

⁷Note, 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 Γ 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).

2.2.5 Implementation in R: The `lavaan` Package

It is further instructive to look at how CFA modeling is implemented and particularly how users can specify model specifications. In doing so, it becomes clearer which information researchers must provide with regard to the model structure. Thus, this subsection contains a brief explanation of the model syntax for CFA models in the R-package `lavaan` (Rosseel, 2012) which is used throughout the empirical sections of this thesis. For the interested reader, a proper introduction for this broad package is available on the package website.⁸

The fundamental way in which model structures are formalized in `lavaan` is with a set of formulas that describe the relationships in the model structure. There are four different types of formulas which are determined by their operator:

- `=~` for defining latent variables
- `~~` for variances and covariances
- `~1` for intercepts
- `~` for regressions.

For our purposes, the first three operators are of particular interest. The first operator, `=~`, is the cornerstone of every CFA model and is used to determine which items load on which latent variables. Items will only be included if they are explicitly specified to load on a given latent variable. More specifically, for a latent variable `eta` and three items `Y1`, `Y2`, and `Y3`, this can be written in `lavaan`-syntax as

```
eta =~ Y1 + Y2 + Y3.
```

Note that this computational notation may be counterintuitive given the fact that in the mathematical notation above, the items are generally viewed as a function of latent variable(s).

With the second operator, `~~`, variance or covariance relationships of any type of variable can be determined. For instance, if the specific variables of the three items are supposed to be modeled as being pairwise correlated, this can be written as

⁸<https://lavaan.ugent.be/tutorial/index.html>

```
Y1 ~~ Y1 + Y2 + Y3
Y2 ~~ Y2 + Y3
Y3 ~~ Y3
.
```

For latent variables, this works analogously. However, it is important to note that `lavaan` models latent variables as pairwise correlated by default. At the same time, the default for the covariance structure of the specific variables is to model only their variances. While it was shown how to add covariances to the default, the removal of a covariance can be achieved by explicitly setting the parameter to zero. Suppose the two latent variables `eta1` and `eta2` should be modeled as being uncorrelated, then the default can be overridden by including in the model formulas

```
eta1 ~~ 0*eta2.
```

The notation used to achieve this can also be used more generally. As mentioned before, it is common to either fix a parameter to a constant or to lump a set of parameters together under a single parameter. Most frequently this is done with the loadings so consider the following two lines

```
eta1 =~ Y1 + Y2 + 2*Y3 + a*Y4
eta2 =~ Y4 + Y5 + a*Y6
.
```

Adding these coefficients to the formulas has two effects. The coefficient `2*` sets the loading of `eta1` on `Y3` to 2 and effectively excludes it from the estimation procedure. The other coefficient `a*` on the other hand “combines” the loadings of `eta1` on `Y4` and the loading of `eta2` on `Y6`. Thus, a single parameter is estimated that is then used for both of these loadings. Note that the name of this new parameter can be arbitrarily set and that it is possible to combine more than two parameters. Further, these parameters don’t even have to belong to the same family of parameters. For example, it is possible to force the variance of a latent variable to be equal to a loading, although it of course becomes more difficult to justify such structural modeling decisions.

Finally, the last relevant operator for our purposes, `~1`, is used to include intercepts. By default, `lavaan` centers the items such that intercepts are implicitly included in the model, but usually not displayed given that they’re not of substantive interest for many. They can be modeled explicitly by including the following formula for at least one item *j*

```
Yj ~1
```

or with the use of an argument of the CFA function.⁹ Note that intercepts can likewise be subjected to constraints by including a coefficient in front of the 1.

With these operators and `lavaan`'s defaults it can be very straightforward to formulate specific model structures. For example, the model measuring political ideology that was described above and is shown in Figure 1 can simply be formalized as

```
eta1 =~ Y1 + Y2 + Y3
eta2 =~ Y4 + Y5 + Y6 .
```

because the defaults with regard to uncorrelated specific variables of the items and correlated latent variables are already in line with the model specification.

2.2.6 Factor Extraction

Up to this point, this section was mostly concerned with fitting EFA and CFA models. However, the initial motivation for conducting CFA was to obtain estimates for the latent variable(s) from a CFA model, particularly if it is used as a measurement model. Recall, that estimation of CFA models works because the likelihood function doesn't depend on the latent variable(s). Latent variables thus need to be estimated from the observed items and the fitted model. As is often the case in factor analysis, multiple options exist (for an overview and a derivation, see Hershberger, 2014). However, for practical purposes, most methods yield comparable results (c.f. Beauducel, 2007). A commonly used method is Thomson's (n.d.) regression method which is also the `lavaan` default method. Given a fitted model with estimates $\hat{\Lambda}$, $\hat{\Phi}$, and $\hat{\Psi}$ as well as the covariance matrix $\hat{\Sigma}$ which they imply, the latent variables $\hat{\eta}$ can be estimated with the regression method via

$$\hat{\eta} = \tilde{Y} \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Phi} \quad (22)$$

where \tilde{Y} is an $n \times p$ matrix of n observations and p centered items such that the resulting matrix of latent variable estimates is of dimension $n \times k$.

⁹ `meanstructure = TRUE`

3 Measurement Invariance

To build some intuition of what measurement invariance (MI) is, how violations of it can arise, and what the implications of 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 (η_1) and a cultural dimension (η_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 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 η_1 , while for group *B*, it loads on η_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.

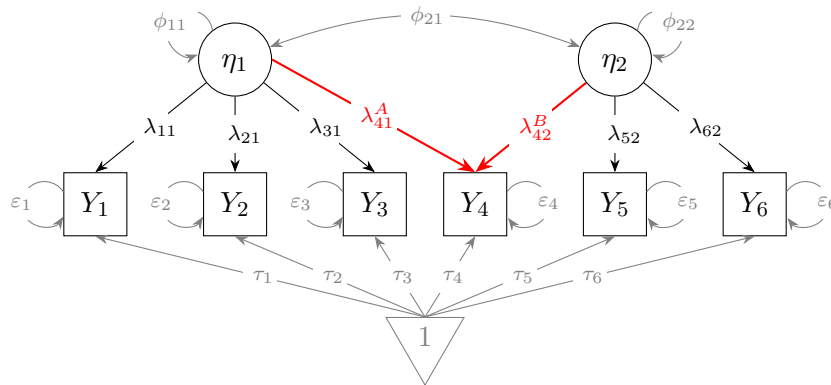


Figure 2: Violation of configural invariance in the model shown in Figure 1.

Less obvious violations could be that certain items 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 and their structural relationships in the model.

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 occur at any sub-population level along any distinguishing line. For instance, measurement non-invariance may arise

from differences due to age groups, gender, etc. As mentioned above, 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).

Formally, MI can be defined as a conditional independence relationship between the items and the latent variable such that \mathbf{Y} is independent of the group $l = 1, \dots, g$ given $\boldsymbol{\eta}$:

$$\mathbf{Y} \perp\!\!\!\perp l \mid \boldsymbol{\eta}. \quad (23)$$

Letting $f(\mathbf{Y})$ denote the distribution of \mathbf{Y} , we can equivalently write

$$f(\mathbf{Y} \mid \boldsymbol{\eta}, l) = f(\mathbf{Y} \mid \boldsymbol{\eta}) \quad (24)$$

by borrowing from Mellenbergh (1989).

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 $l = 1, \dots, g$. This is equivalent to fitting g independent CFA models of the form

$$\mathbf{Y}^l = \boldsymbol{\tau}^l + \boldsymbol{\Lambda}^l \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l, \quad (25)$$

where the superscript index is adapted from Davidov et al. (2014) and should not be confused for a power. In the fitting of a MGCFA model for a given model structure, each free parameter is simply estimated independently for each group.

Continuing in this notation, *configural invariance* assumes that the same loading structure is appropriate across all groups. For configural invariance to hold, the same items must load on the same latent variables 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 $\boldsymbol{\Lambda}$ of the model to hold across groups. Formally, let λ_{ij}^l denote the i^{th}

manifest variable's loading on the j^{th} latent variable in group l . Then, given the proposed structure Λ across all groups, configural invariance is satisfied if

$$\mathbb{1}_{\{\lambda_{ij}^1=0\}} = \mathbb{1}_{\{\lambda_{ij}^2=0\}} = \dots = \mathbb{1}_{\{\lambda_{ij}^g=0\}} \quad \forall i = 1, \dots, p \ \& \ j = 1, \dots, k, \quad (26)$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function. To the attentive reader, it should have appeared that the first violation of MI in the example above as shown in Figure 2 was a textbook violation of configural invariance. Specifically, in the example, configural invariance is violated because

$$\mathbb{1}_{\{\lambda_{41}^A=0\}} = 0 \neq 1 = \mathbb{1}_{\{\lambda_{41}^B=0\}} \ \& \quad (27)$$

$$\mathbb{1}_{\{\lambda_{42}^A=0\}} = 1 \neq 0 = \mathbb{1}_{\{\lambda_{42}^B=0\}}. \quad (28)$$

It should be fairly obvious that the consequences of assuming a common structure Λ 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.

$$\lambda_{ij}^1 = \lambda_{ij}^2 = \dots = \lambda_{ij}^g \quad \forall i = 1, \dots, p \ \& \ j = 1, \dots, k. \quad (29)$$

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 variables outside the model can be compared meaningfully (Davidov et al., 2014; Steenkamp & Baumgartner, 1998). Yet, cross-group comparisons of estimated latent means may still be invalidated by 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

$$\tau_i^1 = \tau_i^2 = \dots = \tau_i^g \quad \forall i = 1, \dots, p. \quad (30)$$

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 when all types of MI hold (c.f. Davidov et al., 2014).

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

$$\Psi^1 = \Psi^2 = \dots = \Psi^g. \quad (31)$$

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. Particularly, the remainder of this thesis focuses on metric and scalar MI while assuming that configural invariance holds. This focus was made for two reasons. First, configural invariance as the lowest level of MI is often supported empirically (Davidov et al., 2014). It thus poses less of a problem for applied researchers. Second, configural invariance is by definition a model property. As will become apparent in the following subsection, metric and scalar MI can also be considered at the item-level at which the various detection methods, discussed and introduced in this thesis, are applicable.

3.2 Partial Measurement Invariance

Thus far, MI has been considered as a model property. However, metric and scalar 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. Given prior research design considerations that the study populations must at least in theory be comparable, it appears much more likely that only a subset of items will function differently across groups. 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.

In the literature on CFA, it appears to be much more common to focus on MI as a model property. At the same time, the closely related literature on item response theory (IRT) makes the distinction between MI at the item versus model level much more frequently. In fact, violations of MI in IRT are usually referred to as differential item functioning (DIF) at the level of individual items and differential test functioning (DTF) at the model level (c.f. Drasgow et al., 2018; Thissen et al., 1993). However, 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, I decided to stick to the conventional labels of MI in the CFA literature even though it is somewhat underdeveloped in this regard.¹⁰

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. This is the primary motivation for writing 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 depends on the context of the study and in particular its research question. Thus, no general recommendations 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

¹⁰Note that there have been attempts to unify the research on MI, DTF, and DIF (e.g. Stark et al., 2006).

a non-invariant item or account for its contribution to biased estimates of latent variables. 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 or account for their bias contribution otherwise.

3.3 Global Test of Measurement Invariance

Before turning to the detection methods at the item level, this subsection briefly introduces the most common tests of global metric and scalar MI. These tests are the most popular choice for assessing global MI in applied research. As such, they are also used in the original study by Castanho Silva et al. (2020), which serves as the basis of the application towards the end of this thesis. This subsection therefore also prepares the reader for the empirical section of this thesis.

At their core, global tests of metric and scalar MI are constructed by fitting several nested MGCFA models and comparing their goodness-of-fit (Jöreskog, 1971). To build some intuition, 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 given its structure is constrained to equality across all groups, effectively turning it into a single CFA model where groups are altogether ignored as in (13). For the second model, fully unconstrained means that each free parameter in the model structure is estimated independently for each group, such that it independently fits the same model to each group as in (25).

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 take the fully unconstrained model in (25) as a baseline model and incrementally impose across-group equality constraints for families of parameters. If the model fit of the baseline model and a model with a constrained family of parameters is sufficiently similar, we can infer invariance of said family of parameters. The underlying intuition is that, on the one hand, the loss in flexibility from these constraints on a given set of parameters does not affect the model's goodness-of-fit significantly if the parameters are sufficiently similar across groups. On the other hand, constraining parameters that violate MI would result in a significant decrease in goodness-of-fit.

In the following, the baseline model is referred to as $\mathcal{M}_{\text{base}}$. As families of parameters, we will firstly consider the loadings for metric invariance and secondly both the loadings and intercepts in the model. For testing global metric invariance, constraining all loadings in

$\mathcal{M}_{\text{base}}$ to equality across groups yields the *weakly constrained model*:

$$\mathcal{M}_{\text{weak}} : \mathbf{Y}^l = \boldsymbol{\tau}^l + \boldsymbol{\Lambda} \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l, \quad (32)$$

where the omission of the group superscript l on $\boldsymbol{\Lambda}$ signifies the equality constraint. For testing scalar invariance, additionally constraining the items in $\mathcal{M}_{\text{weak}}$ yields the *strongly-constrained model*

$$\mathcal{M}_{\text{strong}} : \mathbf{Y}^l = \boldsymbol{\tau} + \boldsymbol{\Lambda} \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l. \quad (33)$$

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

$$\text{Weak MI } H_0 : \Sigma(\mathcal{M}_{\text{weak}}) = \Sigma(\mathcal{M}_{\text{base}}) \quad (34)$$

$$\text{Strong MI } H_0 : \Sigma(\mathcal{M}_{\text{strong}}) = \Sigma(\mathcal{M}_{\text{weak}}) \quad (35)$$

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 Γ , defined in equation (21) for comparing two nested models, we have

$$\Gamma(\mathcal{M}_{\text{base}}, \mathcal{M}_{\text{weak}}) \stackrel{H_0^{\text{weak}}}{\sim} \chi_{(g-1)d_{\text{freeload}}}^2. \quad (36)$$

where, d_{freeload} is the number of free loading parameters in the baseline model such that $(g-1)d_{\text{freeload}}$ 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

$$\Gamma(\mathcal{M}_{\text{weak}}, \mathcal{M}_{\text{strong}}) \stackrel{H_0^{\text{strong}}}{\sim} \chi_{(g-1)p}^2. \quad (37)$$

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. Dras-

gow 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 include the *root mean square error of approximation* (RMSEA; Steiger and Lind, 1980) and the *comparative fit index* (CFI; Bentler, 1990). The CFI is introduced formally in the appendix to this thesis because it is used as the goodness-of-fit measure for one of the existing methods discussed in the next section. 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

As the previous section has shown, MI can be viewed as either a model property or a property of the items in the CFA framework. The latter view enables researchers to address MI issues in their models in several ways to improve the measurement of latent variables. However, which items are invariant and which are not is generally not known. 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 significantly faster than existing methods.

Before turning to the detection methods, the scope of this and the following sections needs to be limited. First, I limit myself to a single factor model, i.e. a simple model with a single latent variable and p items. This greatly simplifies the formalization of these detection methods and is also how they are introduced in the literature. Nonetheless, implementing them for more complex CFA models is fairly straightforward by simply repeating the procedure for each latent variable in the model. I briefly discuss how this can be done in the application section towards the end of the thesis. Second, as mentioned before, MI at the item level is primarily 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. Finally, in a Bayesian setting, there exists another approach for detecting non-invariance at the item level: The general idea is to parametrize the across-group difference for intercepts and loadings in the model (Muthén & Asparouhov, 2013). Obviously, this requires the entire fitting of the CFA model to be conducted via Monte Carlo sampling. Since all other detection methods work in a frequentist setting and because sampling methods are computationally demanding, I omit this approach.

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 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 simultaneously. 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.¹¹ 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 . Further research could explore how these methods can be accommodated for this task.

4.1.1 Janssens (J)

First, another early detection method was devised by Janssens et al. (1995) and 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 which resembles configural invariance. To illustrate, the method wouldn't be able to identify a non-invariant item for which a sign flip existed across two 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, for each item $i = 1, \dots, p$ in 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. Let ϕ_{i1}^l and ϕ_{i2}^l denote the corresponding p-values in group l , for example using a Wald test.¹² For a given significance level α , item i is said to violate scalar invariance if $\phi_{i1}^l < \alpha$ for at least one l , but at the same time $\phi_{i1}^{l'} \geq \alpha$ for at least one l' . The same goes for metric invariance by considering the corresponding p-values. Evaluating violations of

¹¹In a baseline model with a single latent variable, the number of models for comparison would double.

¹²Which is what the `lavaan` package provides for each parameter.

either type of MI, we have as the set of identified items of this approach

$$S_J := \left\{ i \mid \varphi_{i1}^l < \alpha \ \& \ \varphi_{i1}^{l'} \geq \alpha \text{ or } \varphi_{i2}^l < \alpha \ \& \ \varphi_{i2}^{l'} \geq \alpha, \text{ for some } l, l' \right\}. \quad (38)$$

4.1.2 Modification Indices (MInd)

A second and prominent approach is based on modification indices (MInd) (Sörbom, 1989). 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 model. In the context of non-invariant item detection, MInd refer to the additional parameters used when lifting equality constraints for individual items in a strongly 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 indicating non-invariance of the item. Conveniently, what constitutes high MInd can be quantified because the MInd approach can be framed as a LR test 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 strongly constrained MGCFA model and \mathcal{M}_i a model which is identical except for the modification of lifting the equality constraints on the loadings and intercepts of item i . Since, \mathcal{M} is nested in \mathcal{M}_i , the MInd can be written as the LR statistic of these models

$$\Gamma(\mathcal{M}_i, \mathcal{M}) \stackrel{H_0}{\sim} \chi_{2(g-1)}^2. \quad (39)$$

Let t_α be the critical value of the χ^2 -distribution with $2(g-1)$ degrees of freedom at significance level α , then the set of non-invariant items S_{MInd} identified by this method is

$$S_{\text{MInd}} := \left\{ i \mid \Gamma(\mathcal{M}_i, \mathcal{M}) > t_\alpha \right\}, \quad (40)$$

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

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 reference items, i.e. the marker items in CFA, whose loadings are set to 1 for the model to be identified. They argue that

procedures failing to do so may lead to inaccurate results because marker items are effectively constrained to across-group equality as well. To solve this issue, their procedure repeatedly changes the reference item while testing for MI. Non-invariant items are then detected by means of a *triangle heuristic* (Cheung & Rensvold, 1999). Formally, their 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 intercept¹³ 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 with changing reference items $j = 1, \dots, p, j < i$. Taken together, these steps yield a procedure which involves one baseline model and $p(p-1)/2$ models each nested in the baseline model.

Cheung and Rensvold (1999) propose using χ^2 -tests of these nested models, so we can again write the test in terms of the LR-statistic

$$\Gamma_{ij} = \Gamma(\mathcal{M}, \mathcal{M}_{ij}) \stackrel{H_0}{\sim} \chi^2_{2(g-1)} \quad (41)$$

These tests can then be arranged in a strictly lower triangular matrix Γ of test statistics

$$\Gamma := \begin{bmatrix} \Gamma_{21} & & \\ \vdots & \ddots & \\ \Gamma_{p1} & \dots & \Gamma_{p(p-1)} \end{bmatrix}. \quad (42)$$

According to Cheung and Rensvold's (1999) *triangle heuristic*, this matrix can be simultaneously permuted with a suitable permutation matrix P by rows and columns to yield a matrix $\tilde{\Gamma} = P\Gamma P^\top$, maximizing the number of consecutive rows counted from the first row that include no statistic that exceeds the critical value of the $\chi^2_{2(g-1)}$ -distribution for some significance level α . In other words, the permutation should rearrange the items such that significant statistics appear in the lower rows of $\tilde{\Gamma}$. Cheung and Rensvold (1999) then consider those items which appear below the last row that contains no significant test statistics in $\tilde{\Gamma}$ to be non-invariant. More formally, let $\pi(i)$ be the position of item i in the permutation yielding $\tilde{\Gamma}$ and let l_α 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

$$S_{\text{CR}} := \left\{ i \mid \pi(i) \leq l_\alpha \right\}. \quad (43)$$

Note that $\tilde{\Gamma}$ 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 re-

¹³Note, that the original paper remains silent on what to do with intercepts and only talks about constraining the loadings. After some initial testing, I added constraints on the intercepts which improved the method's performance in detecting items violating scalar MI.

sulting sets are valid. They argue that the choice must be “made in light of substantive issues and underlying theory” (Cheung & Rensvold, 1999, p.12). For lack of a better alternative, I arbitrarily choose the first permutation that contains the maximal number of zero-rows in my implementation of their procedure.

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. It is similarly based on model comparisons of goodness-of-fit, just as the CR and MInd approaches. Yet, it differs from the previously described methods in two fundamental ways. First, instead of constraining parameters, it completely removes items from the model one at a time. Second, rather than using a LR-test, the authors propose using the CFI. More specifically, the procedure works by first fitting a strongly constrained MGCFA model \mathcal{M} as the baseline and determining its CFI. Additionally, for each $i = 1, \dots, p$, a model $\mathcal{M}^{(-i)}$ is fitted where item i has been removed from the baseline model, leaving all other model choices untouched. The intuition is that if item i is indeed non-invariant, its deletion from the baseline 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. 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. To summarize, the estimated set of non-invariant items S_{BV} of this procedure is given by

$$S_{BV} := \left\{ i \mid \text{CFI} \left(\mathcal{M}^{(-i)} \right) \geq \text{CFI}(\mathcal{M}) + 0.01 \right\}. \quad (44)$$

4.2 A Novel Approach to Non-invariant Item Detection

[Advantages: Also works if first item is non-invariant without awkward reordering of items (Compare performance for subsets when y1 truly is invariant). Fast.] [Disadvantages:]

The novel approach proposed and tested in this thesis deviates from the existing approaches in several ways. Most fundamentally, instead of being based on model comparison or relying on the model parameters *per se*, it makes direct use of the implications of the relationship between latent variables and items in CFA models. To illustrate the core idea of this novel approach, suppose we have a single latent factor model with p items for the latent variable η . Having fitted such a model, we can obtain estimates for our latent variables $\hat{\eta}$. Because the assumed relationship between η and each item Y_i is linear, we can linearly regress Y_i on $\hat{\eta}$ and obtain residuals from each regression, denoted $\hat{\epsilon}_i$. More

formally, we have

$$\hat{\varepsilon}_i := Y_i - \hat{\mathbb{E}}[Y_i|\hat{\eta}], \quad (45)$$

where $\hat{\mathbb{E}}[Y_i|\hat{\eta}]$ are the fitted values from the regression of Y_i on $\hat{\eta}$. By construction, the residuals of a linear regression model have mean zero and are uncorrelated with their linear predictor when considering all data that were used in the estimation of the regression model. However, these two properties do not necessarily hold for subsets of the data. In fact, they don't generally hold for subgroups in the sample if item i suffers from metric or scalar non-invariance. On the contrary, we would expect the residuals in some or all groups to have non-zero means or for them to be correlated with the latent variable. To further illustrate this, Figure 3 visualizes the differences between residuals from a data-generating process (DGP) that emulates measurement invariance or violations thereof. For all panels the residuals were obtained from a single linear regression that ignores the group structure of the data. In a next step, these residuals are regressed on the latent variable using separate regressions for each group. The fitted lines of these secondary regressions are shown in the figure. Panel A shows the setting where invariance is given, the data come from the same DGP across groups, and the only difference between the groups is in the true latent mean with the red group scoring higher than the black group. 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 η is increased by 0.5 units. Finally, panel D combines both violations using the same values. It is easy to see that for panel A the two criteria of zero mean and vanishing correlation hold in both sub-groups. On the other hand, the remaining panels, reflecting violations of MI, exhibit deviations from at least one of these criteria for at least one of the two groups.

To summarize, the example shows that if MI is violated for a given item, the result is a deviation from these two properties of residuals in the groups. Furthermore, it shows that it's possible to visualize the MI status of the items in CFA models, which may be helpful for applied researchers. The methods required for creating such visualizations are the bread and butter of empirical researchers and could therefore contribute to spreading the use of item-level detection methods for violations of MI. The existing methods, on the other hand, don't have this advantage. Instead, their reliance on ML theory and the use of many submodels may pose a serious hurdle for newcomers to CFA and MI.

The use of residuals of the linear relationships in CFA models for diagnostic purposes is by no means an innovative idea. For example Costner and Schoenberg (1973) use correlations of all items' residuals to identify relationships that are missing from the specified model.¹⁴

¹⁴However, 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)

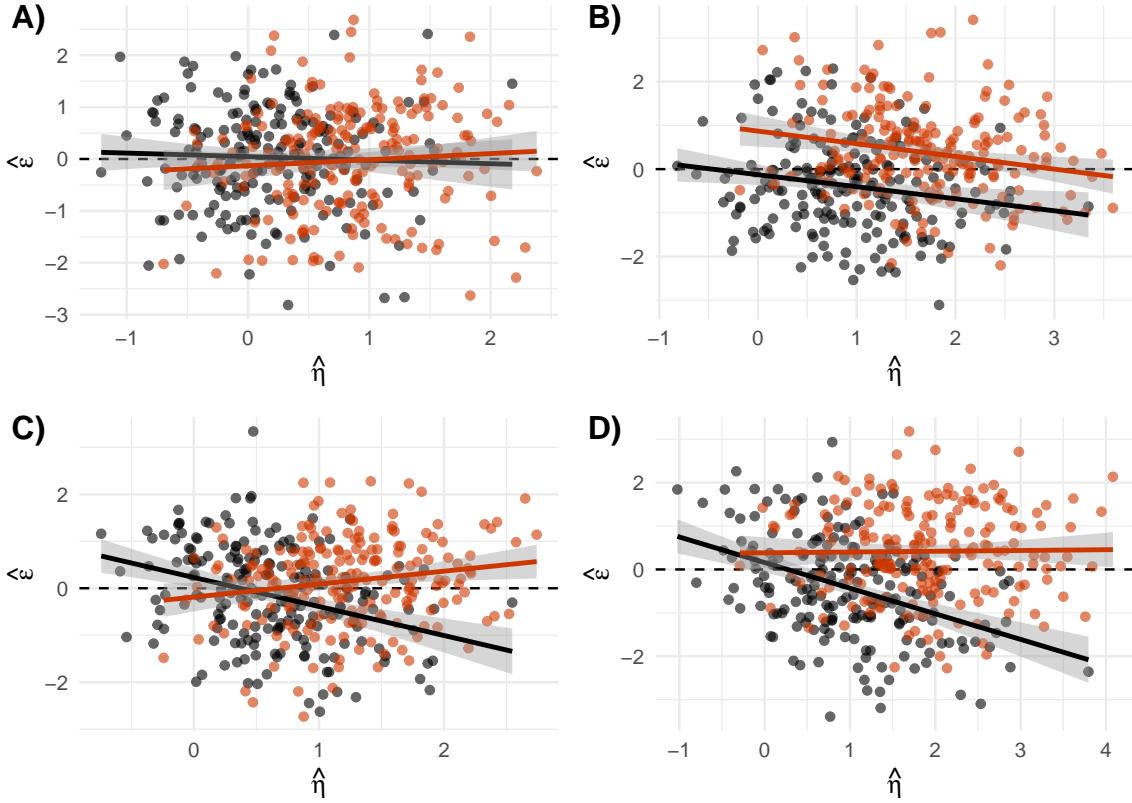


Figure 3: Illustration of the residuals from four DGPs across two groups (black and red).

Regardless, to the best of my knowledge, residual analysis has not been used for detecting non-invariance. The original contribution of this thesis is therefore to devise a method for detecting non-invariant items under partial MI by studying patterns in the residuals of the linear relationships of a given CFA model. In this endeavor, the two components of a) across-group comparisons of the residual means and b) correlations between the residuals and the latent variable within each group must be formalized for a proper hypothesis test. In the following, I show that the first component can be formulated as a standard one-way analysis of variance (ANOVA) and the second component can be studied by means of a coefficient test which is equivalent to a correlation test in the absence of control variables. I then describe how to aggregate these two components to yield a test of non-invariance at the item level. Finally, I argue that a step-wise version of the novel method may further improve its performance and describe how that version is implemented.

4.2.1 ANOVA Component

Recall that for the ANOVA part, the implication of an invariant item is that the residuals have the same expectation across all groups. Let

$$\mu_i^l := \mathbb{E} \left[\hat{\epsilon}_i^l \right] \quad (46)$$

denote the expected residual for item i in group l . For each item, we can then write the null hypothesis for the ANOVA component of the procedure as the global null hypothesis of a single-mean model, i.e.

$$H_0 : \mu_i^1 = \dots = \mu_i^g, \quad (47)$$

which can be conducted with an F -test. More specifically, the test statistic for item i with $N = n^1 + \dots + n^g$ samples is given by the ratio of treatment and error mean squares:

$$F_i := \frac{MST_i}{MSE_i} = \frac{\frac{1}{(g-1)} \sum_{l=1}^g n^l (\hat{\mu}_i^l - \hat{\mu}_i)^2}{\frac{1}{(N-g)} \sum_{l=1}^g \sum_{j=1}^{n^l} (\hat{\varepsilon}_{ij}^l - \hat{\mu}_i^l)^2} \stackrel{H_0}{\sim} F_{(g-1), (N-g)}, \quad (48)$$

where $\hat{\mu}_i^l$ and $\hat{\mu}_i$ are the sample means of the residuals in group l and in the full sample of size N , respectively.¹⁵ For a more detailed account, refer to an introductory ANOVA book, e.g. Oehlert (2000).

4.2.2 Correlation Component

With regard to the correlation between the residuals and the latent variable, we can test the linear relationship by testing the corresponding coefficients in separate regressions of $\hat{\eta}^l$ on $\hat{\varepsilon}_i^l$ for each item and each group.¹⁶ Let β_i^l denote the coefficient of $\hat{\eta}^l$. The null hypothesis of no correlation can then be written as

$$H_0 : \beta_i^l = 0. \quad (49)$$

More specifically, we have for each item and group the following test statistic:

$$\frac{\hat{\beta}_i^l}{\text{se}(\hat{\beta}_i^l)} \stackrel{H_0}{\sim} t_{n^l-2} \quad (50)$$

and an estimator for the standard error of the coefficient is given by

$$\hat{\text{se}}(\hat{\beta}_i^l) = \sqrt{\frac{\sum_{j=1}^{n^l} (\hat{\varepsilon}_{ij}^l - (\hat{\gamma}_i^l + \hat{\beta}_i^l \hat{\eta}_j^l))^2}{(n^l - 2) \sum_{j=1}^{n^l} (\hat{\eta}_j^l - 1/n^l \sum_{j=1}^{n^l} \hat{\eta}_j^l)^2}}, \quad (51)$$

where $\hat{\gamma}_i^l$ is the estimated intercept of the corresponding regression. Again, details can be found in any standard introductory statistics textbook introducing linear regression, e.g. Fahrmeir et al. (2013).

¹⁵Note, that the latter sample means is zero by construction. However, it is kept in equation (48) to emphasize that the statistic is the standard ANOVA F -test.

¹⁶Equivalently, this can of course be done in a single regression where $\hat{\eta}$ is interacted with the group.

4.2.3 Aggregating the Components

What remains to be done is to aggregate the two components to a single test at the item level. To this end, first note that for each item, the procedure is comprised of $g + 1$ individual tests: One global ANOVA test for the residual means and g correlation tests. Let \wp_{iu} denote the p-value of the u^{th} test for item i . These p-values can be aggregated by simply considering the minimal p-value for each item and applying a Bonferroni correction such that

$$\tilde{\wp}_i := (g + 1) \min(\wp_{i,1}, \dots, \wp_{i,g+1}), \quad (52)$$

is a p-value for the test of item i 's non-invariance. In a next step, a Holm-Bonferroni correction at the level of items can be applied to yield the set of identified non-invariant items for which the item-level test is rejected:

$$S_{R1} := \left\{ i \mid (p - \pi(i) + 1) \tilde{\wp}_i < \alpha \right\}, \quad (53)$$

where $\pi(i)$ denotes the position of the i^{th} item when arranging the p-values in ascending order.

4.2.4 Step-wise Version

Perhaps the clearest drawback of this approach, is that it hinges on $\hat{\eta}$ being reasonably close to the true latent variable. If MI is violated strongly, i.e. for many items then $\hat{\eta}$ has little resemblance with the true latent variable even if the model structure, i.e. the relationships between latent variables and items is correct. Thus, this, but also the other existing approaches would fail to correctly distinguish non-invariant items. However, the degree of non-invariance at which this and existing methods still work can only be studied with simulations. One potential way of ameliorating this issue is to implement a step-wise version of this approach. Starting with a given CFA model \mathcal{M} , the step-wise approach first selects the item with the lowest p-value in the R1 approach. If its p-value is below the given significance level α , it is removed from the model entirely and the next iteration of the process begins. The motivation for doing so is to incrementally improve the estimates of the latent variable by refitting the model with the worst item removed in each iteration until no more items are detected as being non-invariant for a given α -level. Note, that this idea of improving the model by removing items entirely has some resemblance with the approach by Byrne and Van de Vijver (2010). Further note that the fundamental idea of the original method *R1* still applies. In fact, computationally, its implementation can simply be reused in each iteration of the step-wise approach. Instead of using the set-builder notation, the set of non-invariant items in a CFA model \mathcal{M} can best be described by showing the following algorithm.

1. Set $t = 0, S_{R2} = \{\}$
2. Apply R1 to the initial model $\mathcal{M}_0 = \mathcal{M}$, yielding $\tilde{\varphi}$
3. while $((p - t) \min \tilde{\varphi} < \alpha)$ {
 - (a) $S_{R2} = S_{R2} \cup \arg \min_i \tilde{\varphi}_i$
 - (b) Update model $\mathcal{M}_{(t+1)} = \mathcal{M}_t^{(-\arg \min_i \tilde{\varphi}_i)}$, removing item $\arg \min_i \tilde{\varphi}_i$
 - (c) Apply R1 to \mathcal{M}_{t+1} and update $\tilde{\varphi}$
 - (d) $t = t + 1$

4.3 Implementation

All detection methods were implemented in the R programming language (R Core Team, 2020; v4.0.2) and are publicly available in the GitHub repository for this thesis.¹⁷ Every step within the CFA framework, i.e. model fitting, testing, etc., was done using the `lavaan` package (Rosseel, 2012; v0.6-9).

4.4 Overview over Methods for Detecting Non-invariant items

¹⁷<https://github.com/pitrieger/masterthesis/tree/main/Rscripts/simulation> for models with a single latent variable and <https://github.com/pitrieger/masterthesis/tree/main/Rscripts/application> for models with multiple latent variables (see section 6 for further details.)

Method	Type	Reference	Summary
J	Parameter inspection	Janssens et al. (1995)	Identifies items by comparing statistical significance of intercepts and loadings across groups. If they are significant in some groups, but not others, the corresponding item is considered non-invariant.
MInd	Model comparison (χ^2)	Sörbom (1989)	Item-wise lifting of equality constraints across groups. Decision based on comparison with strongly constrained model with LR test.
CR	Model comparison (χ^2)	Cheung and Rensvold (1999)	Item-wise imposing of equality constraints across groups. Decision based on comparison with fully unconstrained model with LR test. Additionally takes into account marker method by systematically varying reference item.
BV	Model comparison (CFI)	Byrne and Van de Vijver (2010)	Entirely removes one item at a time from the model. Decision based on difference in CFI compared to strongly constrained model.
R1	Residuals	<i>original</i>	Obtains residuals from regressing estimated latent variables on single items. For each item, tests equal means of residuals across groups and tests for vanishing correlation between residuals and items within each group. Items classified as non-invariant if means are different across across groups or if there is non-vanishing correlation in at least one group.
R2	Residuals	<i>original</i>	Like R1, but instead of identifying non-invariant items simultaneously, works iteratively and removes one item at a time.

Table 2: Overview of different detection methods.

5 Simulation Study

This section provides a comparison of the various detection methods by means of a simulation study. The obvious benefit of a simulation study is that an evaluation of the different methods is possible because there is an objective and known truth. However, it should be noted that simulation studies come with the obvious caveat that their external validity is limited.

I begin with a description of the data generating process that is used for the simulation study. Then, I introduce the implementation and general setup and discuss how the results will be analyzed. Finally, I present and interpret the detailed results.

5.1 Data Generation

For simulating data, I rely on the data generating process under partial non-invariance laid out by Pokropek et al. (2019). It generates data from a single-latent variable model for several groups under different settings of partial non-invariance. Most of the fixed simulation parameter values were also taken from Pokropek et al.'s (2019) original simulation study. I consider all possible combinations of the different parameter values summarized in Table 3 with the exception of combinations that would have more non-invariant than invariant items and those where the number of affected groups doesn't yield a natural number (see details below). In total, this yields 2016 different simulation settings. In the following, the steps of the data generating process are summarized.

For each group $l = 1, \dots, g$, a group-specific mean and standard deviation of the latent variable are generated. The g true latent means μ are obtained from a normal distribution with mean zero and a standard deviation of 0.3, i.e.

$$\mu_l \sim \mathcal{N}(0, 0.3). \quad (54)$$

The g true standard deviations σ are the absolute value of normal distribution draws with mean one and a standard deviation of 0.1, i.e.

$$\sigma_l \sim \mathcal{N}^+(1, 0.1). \quad (55)$$

These parameter samples are then used to sample positions on the latent variable η_l for each observation $i = 1, \dots, n$ within each group. Note that n is the number of observations in each group such that in total there are $N = gn$ observations with all groups having equal size. Therefore, the observation index i is nested in the group index l , denoted $l(i)$. The latent variables are then sampled as

$$\eta_{l(i)} \sim \mathcal{N}(\mu_l, \sigma_l). \quad (56)$$

Turning to the items, intercepts τ and loadings λ for each of the $j = 1, \dots, p$ items Y_j are sampled. The loadings come from a normal distribution with mean 0 and standard deviation of 0.5, i.e.

$$\tau_j \sim \mathcal{N}(0, 0.5) \quad (57)$$

and the loadings are generated from a uniform distribution on $[0.65, 0.85]$, i.e.

$$\lambda_j \sim \text{Unif}(0.65, 0.85). \quad (58)$$

Scores $Y_{l(i)j}$ for each observation and item in a given group are finally sampled from a normal distribution with mean $\tau_j + \lambda_j \eta_{l(i)}$ and standard deviation $1 - \lambda_j^2$, i.e.

$$Y_{l(i)j} \sim \mathcal{N}(\tau_j + \lambda_j \eta_{l(i)}, 1 - \lambda_j^2) \quad (59)$$

Note that if it stopped here, the data generating process would reflect perfect MI. To create a setting of partial MI, a total of hg groups and k items is randomly selected, where $h \in \{0.25, 0.5\}$. These selected groups and items will constitute the origin of non-invariance in the data. To introduce non-invariance, the previously sampled intercepts and loadings for these affected groups and items are altered with a bias of magnitudes δ_1 and δ_2 , respectively, where $\delta_1, \delta_2 \in \{0, 0.2, 0.4\}$. As a result, there are four different settings with regard to the type of (violation of) MI: 1) metric and scalar non-invariance ($\delta_1 > 0$ & $\delta_2 > 0$), 2) purely scalar non-invariance ($\delta_1 > 0$ & $\delta_2 = 0$), 3) purely metric non-invariance ($\delta_1 = 0$ & $\delta_2 > 0$), and 4) perfect MI ($\delta_1 = \delta_2 = 0$). Additionally, the sign of each bias is randomly sampled for each group and item. Let l' and j' denote a group and an item that were sampled to be affected by non-invariance. Then

$$Y_{l'(i)j'} \sim \mathcal{N}(\tau_{j'} \pm \delta_1 + (\lambda_{j'} \pm \delta_2) \eta_{l'(i)}, 1 - (\lambda_{j'} \pm \delta_2)^2). \quad (60)$$

Finally, all items are discretized to integer values ranging from -2 to 2 , using ± 0.47 and ± 1.3 as breaks, resulting in a 5-point scale. This step reflects the fact that response scales in survey research are almost exclusively discrete scales with 5-point scales being a very popular choice (Pokropek et al., 2019, c.f.). Nonetheless, these categorical items are treated as continuous, a practice which was shown to be valid for ML estimation by multiple studies (e.g Johnson & Creech, 1983; Muthén & Kaplan, 1985).

Parameter		Values	Comment
Number of observations	n	{100, 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}	
Magnitude of bias on intercepts	δ_1	{0, 0.2, 0.4}	sign of bias randomly sampled for each group and indicator
Magnitude of bias on loadings	δ_2	{0, 0.2, 0.4}	————— —————

Table 3: Simulation parameters.

5.2 Simulation Setup

For each of the 2016 unique simulation parameter value combinations, I simulate 100 datasets according to the procedure introduced above, resulting in 201,600 total iterations. In each iteration, all detection methods are employed to detect non-invariant items. This is done without fine-tuning them with regard to the type of potential violations of MI to mirror the real-world setting where this is generally not known. In other words, detection methods are always set to detect both metric and scalar non-invariant items irrespective of which bias was actually simulated. Each method then returns a set of items that it classifies as non-invariant.

Given the nature of the output of each method and that the truly non-invariant items are known, 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 detection 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: While 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 for the simulation study 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).

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (61)$$

Nonetheless, a good detection method should also have relatively few falsely identified items. While being less problematic, falsely identified items can still be very costly. Additionally, it would be easy to maximize sensitivity with a method that always returns positive prediction. Thus, the methods' performances are also assessed with a 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):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \quad (62)$$

Only methods that have both a high sensitivity and specificity can be considered to perform well in detecting non-invariant items. On the other hand, highly sensitive methods with low specificity classify too many items as non-invariant while highly specific methods with low sensitivity classify too few.

5.3 Results

The remainder of this section presents the key 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, an overview across all simulation settings is provided in Table 5. It includes the aggregate sensitivity and specificity under the presence of non-invariant items with regard to both metric and scalar MI ($\delta_1 > 0$ & $\delta_2 > 0$), scalar MI ($\delta_1 > 0$ & $\delta_2 = 0$), and metric MI ($\delta_1 = 0$ & $\delta_2 > 0$) as well as in the absence of any non-invariant items ($\delta_1 = \delta_2 = 0$), while disregarding the magnitude of δ_1 and δ_2 . For the more detailed analysis below, I focus on the case of simultaneous metric and scalar non-invariance, i.e. where both the intercepts and loading are affected.

Starting with the case of simultaneous metric and scalar non-invariance in Table 5, all methods but the J approach perform well in terms of sensitivity. All other methods detected at least 82% of the items that were truly non-invariant. The highest sensitivity is achieved by the MInd approach, with less than 2% of non-invariant items that were not detected. Yet, this high sensitivity comes at the cost of the lowest specificity across all methods. In other words, the MInd approach has a high number of false positives, classifying too many items as non-invariant. In general, performance in terms of specificity isn't very good with the exception of the BV and R2 approaches which achieve a specificity of around 0.8. While the BV approach exhibits the highest specificity, it isn't as sensitive as the others. In terms of the trade-off between the two metrics, the R2 approach performs best with a decently high sensitivity and the second best specificity. When comparing it to the

Method	$\delta_1 > 0$				$\delta_1 = 0$		
	$\delta_2 > 0$		$\delta_2 = 0$		$\delta_2 > 0$	$\delta_2 = 0$	
	metric & scalar		scalar		metric	none	
	Sens	Spec	Sens	Spec	Sens	Spec	Spec
J	0.518	0.575	0.495	0.582	0.424	0.590	0.595
MInd	0.989	0.307	0.966	0.371	0.936	0.444	0.857
CR	0.833	0.362	0.820	0.364	0.790	0.363	0.401
BV	0.821	0.842	0.649	0.890	0.506	0.881	0.990
R1	0.955	0.482	0.940	0.473	0.798	0.659	0.915
R2	0.935	0.779	0.911	0.780	0.748	0.810	0.937

Table 5: Aggregate performance of the detection methods across all simulation settings. Each performance metric was computed from the classification of at least 100,000 items, but varies due to the $>$ -formulation of item and loading bias. Note that the sensitivity for the case of no bias on either intercepts or loadings is trivially zero and was thus omitted from the table.

R1 approach, it further appears that the step-wise approach yields a substantial increase in specificity for a negligible decrease in sensitivity.

Under scalar non-invariance, a similar picture arises. In general, all methods have a slightly lower sensitivity and a similar or higher specificity. However, the overall ranking of the methods in terms of their sensitivity and specificity remains intact. Thus, the R2 approach still fares best when taking both sensitivity and specificity into account. A notable exception to these similarities is the sensitivity of the BV approach which exhibits a sizeable decrease of almost 0.2 compared to the simultaneous non-invariance case.

For the metric non-invariance setting, the results seem to suggest that this is the most difficult case for all methods. In general, all methods perform even worse than in the pure scalar non-invariance case. Particularly the BV and both R approaches have difficulty detecting violations when only the loadings are affected. Despite the decrease in sensitivity of about 0.18, the R2 approach remains the best performing method under the trade-off between sensitivity and specificity. It still correctly classified 75% of the truly non-invariant items and about 80% of truly invariant items.

In the perfect MI case, i.e. when $\delta_1 = \delta_2 = 0$, all methods but the J and CR approaches fare decently in classifying the items as negatives. Note that the sensitivity in this special case is zero for all methods by construction because there simply aren't any TP cases. Again, the BV approach has the highest specificity, followed closely by the two R methods and the MInd procedure. The relatively high specificity of most methods is encouraging, because it shows that in the case where there truly are no invariance violations, superfluous detection is very rare. There really is no reason not to use these detection methods given that in the best case scenario where they aren't needed, they come at virtually no cost. On the other hand, this reasoning suggests that the widespread use of the J, CR, and MInd approaches may not be advisable.

Before moving on to the more detailed analysis, the J approach requires a small disclaimer.

In contrast to all other methods, it is the only one that consistently performs poorly in terms of both sensitivity and specificity across all types of MI violations. The simulation results therefore corroborate the theoretical weaknesses of the J method, discussed above, and show that it cannot be considered a useful method for detecting non-invariant items under any setting. For completeness' sake, I keep the results for the J approach in all tables and figures below, but mostly disregard it in the interpretation of the results.

5.3.1 Sensitivity

Figure 4 sheds some more light on how sensitivity is affected by distinguishing between the number of observations per group n , the number of items p , and the number of non-invariant items k . Recall that in the following, all comparisons and figures refer to the case where both the intercepts and loadings are simultaneously non-invariant, i.e. the case where $\delta_1 > 0$ & $\delta_2 > 0$. All other simulation parameters were disregarded, so the overall level of sensitivity presented below reflects an average of their parameter values. Further note that the confidence intervals in this and the remaining figures are vanishingly small because of the high number of observations from the simulation settings.

Generally, it can be said that with the exception of the J and the CR method, the sensitivity of all other approaches increases in n . This is especially true at the lower end of group sizes, i.e. when moving from 100 to 200 observations per group. The remaining methods also tend to achieve a slightly higher sensitivity for models with more items and slightly lower sensitivity as the number of affected items increases. The BV approach seems to be affected most by increases in k and also appears to no longer benefit from increases in n , as is shown by the virtually horizontal pink lines in the second and third row. For the three methods with consistently high sensitivity, i.e. MInd, R1, and R2, there is no practically relevant difference anymore when n and p are sufficiently high: If n is at least 500 and p greater than 3, all of these methods are able to detect nearly 100% of truly non-invariant items.

When dissecting the sensitivity by the number of groups and the share of groups that are affected by non-invariance, a similar picture arises. Figure 5 shows that, with the exception of the CR approach, all methods exhibit higher sensitivity as g increases. Particularly for four or fewer groups, performance is sub-par while sensitivity is well over 0.95 when $g = 16$. Comparing the left with the right panel, sensitivity is generally higher when the number of affected and unaffected groups is balanced, i.e. when $h = 0.5$. As a note of caution when comparing the two panels, note that the left panel doesn't include any cases for $g = 2$ because g times h isn't a natural number. The only exception to this generally positive change is the J approach. Furthermore, the BV method benefits most from this balance. To illustrate, the sensitivity for the BV approach increases by almost 0.3 when two instead of just one of four groups are affected by non-invariance.

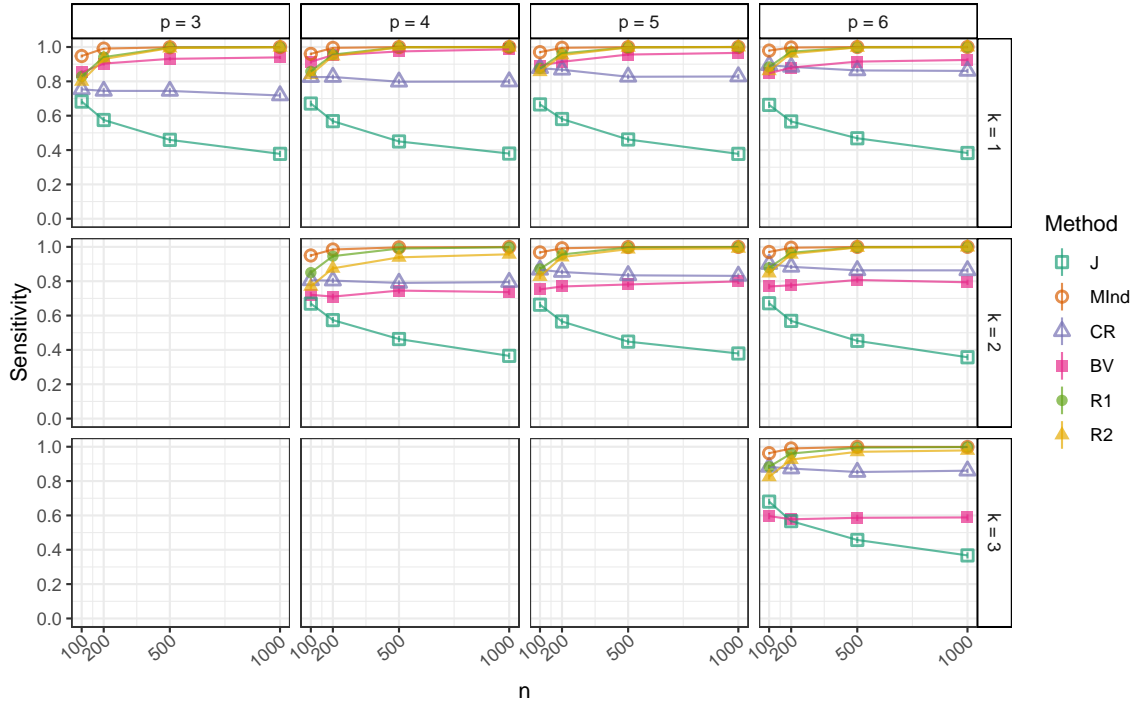


Figure 4: Sensitivity of different detection methods as a function of n , p , and k under partial metric and scalar non-invariance. Vertical lines represent 95% Clopper-Pearson confidence intervals.

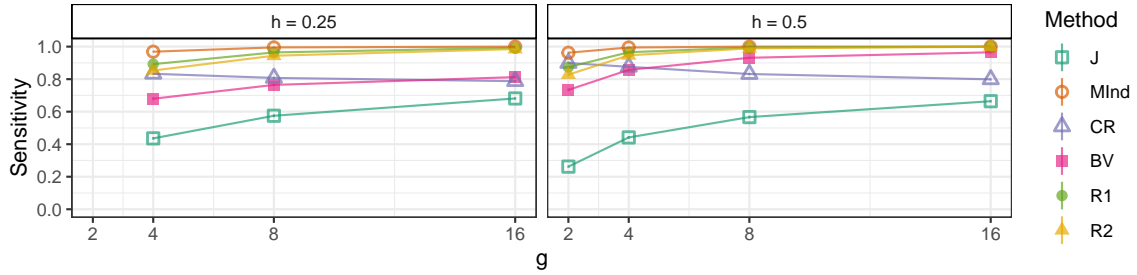


Figure 5: Sensitivity of different detection methods as a function of g and h . Vertical lines represent 95% Clopper-Pearson confidence intervals.

5.3.2 Specificity

As mentioned above, sensitivity alone doesn't allow for an evaluation of the performance of the different methods. By additionally analyzing the specificity of the methods under different simulation settings, we get a much more holistic picture of their performance. Figures 6 and 7 were created analogously to Figures 4 and 5, but plot the specificity instead of sensitivity. A first glance already reveals that there is much more variation compared to sensitivity, both within and across methods. Figure 6 shows that the vast majority of methods tends to detect less of the true negatives as n increases. Put differently, the number of false rejections of the null hypothesis of non-invariance increases with n . However, not all methods are affected in the same way by this. As long as there are more than three items, the best results by far are obtained by the BV method which is also robust to vary-

ing sample size. The poor performance of the BV approach in the top-left panel can be explained by the fact that its item deletion leads to an underidentified model which for computational reasons results in almost all items being classified as non-invariant. However, in the remaining panels of the top row, it consistently achieves a specificity of almost 1. Only the R approaches are able to even come close to this level of specificity. Of the two, the R1 approach is very much affected by increases in n while the R2 approach is fairly consistent. For the MInd and R1 approaches, the results indicate very poor performance and show that they are particularly susceptible to changes in sample size: For the largest sample size of 1000, specificity of the MInd method is as low as 0, meaning that it effectively classifies every item as being non-invariant. As the aggregated results have shown, the CR approach performs relatively poorly in terms of specificity. However, it also seems to be less affected by sample size.

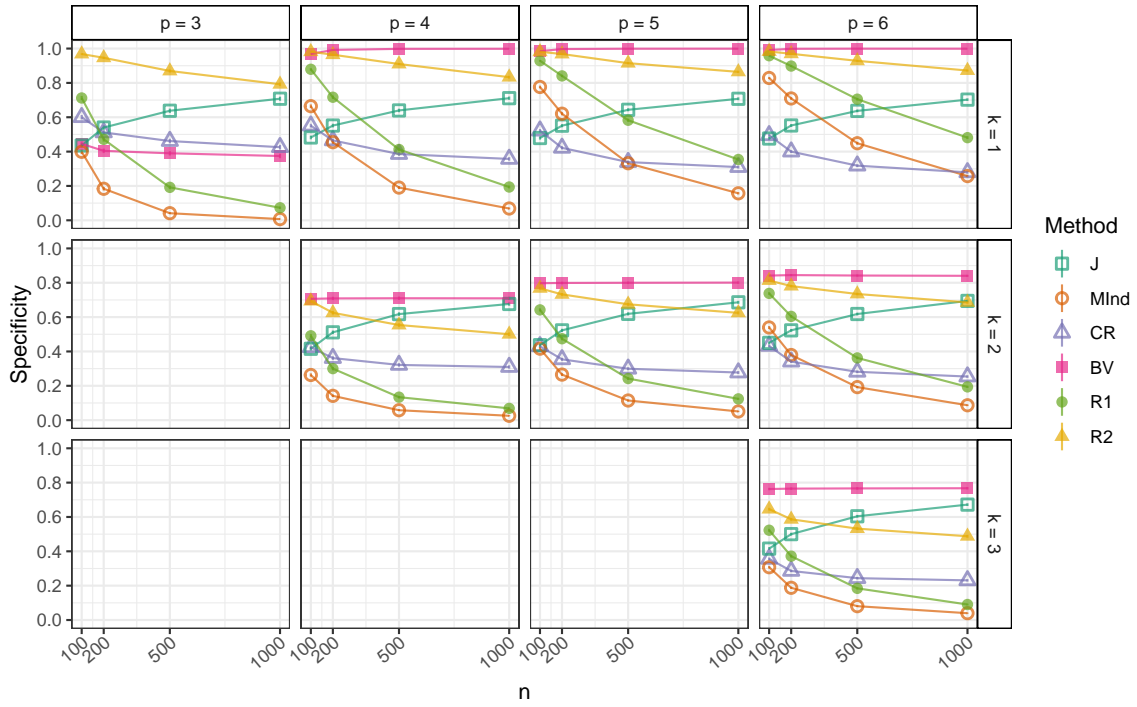


Figure 6: Specificity of different detection methods as a function of n , p , and k . Vertical lines represent 95% Clopper-Pearson confidence intervals.

The same applies to increases in the number of groups, as Figure 7 shows: As g increases, specificity tends to decrease for all but the BV method. Likewise, the R2 approach exhibits an almost horizontal line, indicating that its specificity remains relatively constant for the different settings of g . The effect of the share of affected groups h on the various methods is less unequivocal. This can be seen by moving from the left to the right panel in Figure 7. However, when comparing the two panels, again note that it doesn't include any cases for $g = 2$ for reasons discussed above. In general, the BV, MInd, and R1 approaches perform (slightly) worse in the setting where the number of affected groups is equal to the number of unaffected groups. On the other hand, the CR, R2, and J achieve (slightly) higher specificity in the balanced setting.

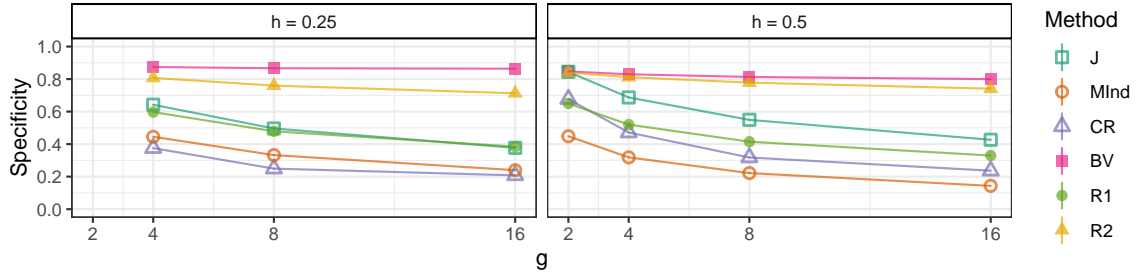


Figure 7: Specificity of different detection methods as a function of g and h . Vertical lines represent 95% Clopper-Pearson confidence intervals.

5.3.3 Specificity under perfect MI

Finally, the various methods' specificity in the case of perfect MI is shown in Figure 8. Recall that sensitivity is trivial in this setting because by construction there are no (true) positive cases. The key finding in the aggregate results above was that specificity is relatively high in the absence of non-invariance. Figure 8 corroborates this finding and additionally shows that this also holds for larger sample sizes. Although specificity decreases for increases in n , this effect of the sample size is much slighter compared to the decreases shown in the corresponding Figure 6. In other words, susceptibility to increases in n is lower than in the presence of non-invariance. When we ignore the overall poorly performing J and CR methods, about 80% or more of all items are correctly classified as being invariant. Here again, the BV method proves to be the most specific method while remaining robust to changes of n . The remaining methods rank similarly compared to the overall specificity discussed in the previous section. Yet, differences between methods are less pronounced in the setting of perfect MI.

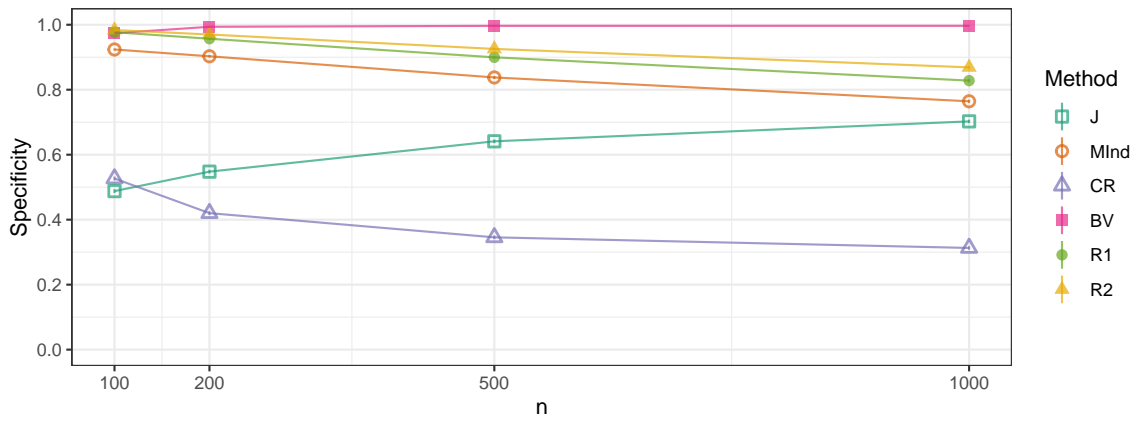


Figure 8: Specificity of different detection methods as a function of n . Vertical lines represent 95% Clopper-Pearson confidence intervals.

5.3.4 Summary

The findings shown above have several implications for the use of these detection methods. In general, some methods seem to work better than others. While the J, MInd, and CR approach cannot be recommended, the remaining three methods, i.e. the BV, R1, and R2 approach, can be recommended generally or at least for certain settings. Broadly speaking, the former group exhibits too many weaknesses, either theoretically or in terms of their sensitivity and specificity. At the same time, the latter three approaches perform decently well in terms of both sensitivity and specificity in almost all setting. In the following, I briefly review each method and discuss under which setting it can be recommended for the detection of non-invariant items.

Starting with the J method, it is unequivocal that this method shouldn't be used at all. Its theoretical weaknesses result in very poor performance in terms of sensitivity because few violations of MI are of the form that can be detected by the J method. In terms of specificity, it fares decently for larger n , where it of course also has lower sensitivity. The MInd approach is plagued by a different issue: While having outstanding sensitivity even under pure metric MI violations, where other methods struggled, its sensitivity is very low. This is the result of the method classifying almost every item as being non-invariant. While this inflates its sensitivity, it does so at the cost of specificity. The method is therefore of little use for applied research. The CR approach similarly has very low specificity and somewhat lower, but very consistent sensitivity across all different types of MI violations and simulation settings. It is easily outperformed by the methods below and can therefore not be recommended.

For the remaining three methods, the performance in terms of both sensitivity and specificity is much better in all or at least some simulation settings. The BV approach can be recommended for almost all settings with the exception of models with only three items. In general, it has particularly high and robust specificity and is reasonably sensitive as long as the ratio of truly non-invariant items to the total number of items isn't too high. Thus, it may be advisable for CFA models consisting of a fairly high number of items and where unnecessary removal of items should be minimized, for instance because it is particularly costly. In all other cases, the R2 approach appears superior to the BV method. Of all contenders, the R2 approach can be recommended most widely. For the R1 approach, specificity is usually too low, but it can still be used as long as the number of observations is relatively low. However, the slight increase in sensitivity compared to the R2 approach is arguably not worth the comparably larger decrease in specificity. For larger sample sizes, it suffers from the same issues as the MInd and CR approaches. The R2 approach on the other hand can be recommended across the board especially when valuing sensitivity over specificity, as is often the case.

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

In this section, the methods for detecting non-invariant items are applied to real-world data in the field of political science. More specifically, I use replication data from a survey fielded by Castanho Silva et al. (2020) who compare several CFA measurement models for populist attitudes. Of course, there is no ground truth against which the results of the detection methods in this section can be compared. Instead, the purpose of this section is twofold. First, it requires an implementation of the methods beyond single-factor models that were used in the simulation study. It thus serves as a proof of concept that the methods can also be used for these more complex models. Second, it contributes to the empirical study of populist attitudes in social science research. While model development is oftentimes theory-driven, Castanho Silva et al. (2020) mention that researchers also take empirical considerations into account. Yet, they seem to focus mostly on loading magnitude rather than the MI properties of their items even if they are ultimately interested in cross-country comparisons. This application thus also serves as an example of how the detection methods can be used for the purpose of model development and improvement.

I begin by briefly summarizing the original paper by Castanho Silva et al. (2020) to make readers familiar with the concept of populism and to highlight the authors findings with regard to global measurement invariance. I then describe idiosyncrasies in implementing the detection methods to accommodate the more complex models. Finally, I present and discuss the results of applying the detection methods to the various models of populist attitudes.

6.1 Synopsis of the Original Paper Castanho Silva et al. (2020)

Castanho Silva et al. (2020) compare seven existing measurement models of populist attitudes using original survey data from nine countries¹⁸ in a CFA framework. More specifically, in their comparison, the authors consider several properties relating to the models' internal coherence, cross-national validity, conceptual breadth, and external validity. Crucially, each respondents was subjected to every question that is required by any of the models so that the models can be fitted using identical samples, which makes their comparison more straightforward. The different contender models were taken from the fast-growing literature on populist attitudes and differ both in terms of their model structure and the survey questions they use as items. Specifically, the authors consider contributions by Akkerman et al. (2014), Castanho Silva et al. (2018), Elchardus and Spruyt (2016), Hobolt et al. (2016), Oliver and Rahn (2016), Schulz et al. (2018), and Stanley (2011).¹⁹ All of these studies build on Mudde's (2004) definition of populism as "a thin-centered

¹⁸Brazil, France, Greece, Ireland, Italy, Mexico, Spain, United Kingdom, and the United States.

¹⁹An overview of all these models and items is available in the supplementary material of the original study: <https://journals.sagepub.com/doi/suppl/10.1177/1065912919833176>

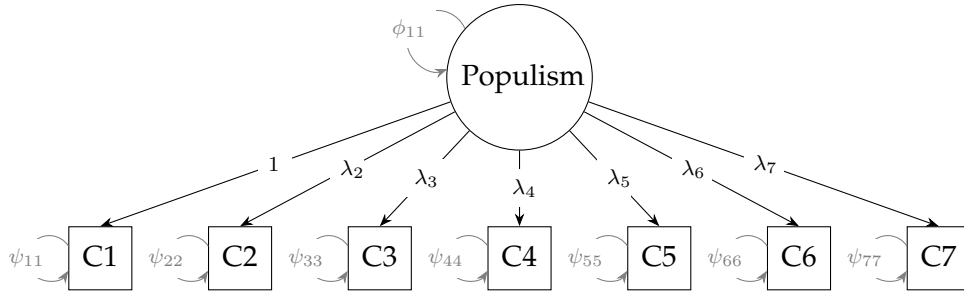


Figure 9: Graphical depiction of Hobolt et al.'s (2016) populism model. Note that intercepts were omitted in this visualization.

ideology according to which society is divided into two homogeneous and antagonistic groups: the 'good people' and the 'corrupt elites'" (Castanho Silva et al., 2020, p. 409-410) and devise a measurement model for this latent concept. People with populist attitudes thus subscribe to the belief that there is a common will of "the people" and that "corrupt elites", or the "establishment", have their own agenda irrespective of the will of the people. There is considerable variation in how the different models go about measuring such populist attitudes. To illustrate this, I focus on the models by Hobolt et al. (2016) and Castanho Silva et al. (2018) which will also be the focus of the empirical analysis of their items' MI properties. They can be viewed as polar opposite examples among the measurement models: While Hobolt et al.'s (2016) model is a very simple single-factor model, the model by Castanho Silva et al. (2018) is much more complex with multiple latent-variables and complex structural features such as equality constraints on some of its loadings. The two models are visualized in Figures 9 and 10, respectively, where intercepts are excluded in the visualizations, but not the models, for the sake of clarity.

Figure 9 illustrates Hobolt et al.'s (2016) model which is very straightforward: It incorporates a single latent variable, i.e. populism, and contains seven items $C1, \dots, C7$ corresponding to seven survey questions. For each of questions, respondents are asked to indicate how much they agree with on a scale from 1 (disagree) to 5 (agree) with the statements shown in Table 6. Furthermore, all items are assumed to load on the single latent factor and to be uncorrelated otherwise. Put differently, when fitting the model, each entry of Λ is estimated freely with the exception of a marker variable while the off-diagonal entries of Ψ are constrained to zero. In essence, the model specification yields a simple EFA model that is rendered identifiable by the marker variable instead of the constraints for an unrotated solution. Their model thus assumes that populism is a single latent variable that directly explains the shared covariance of the answers to these survey questions.

Castanho Silva et al.'s (2018) model on the other hand takes a very different approach which results in a much more complex model. Their starting point is to argue that populism is a multidimensional concept, consisting of *anti-elitism* (ANT), *people-centrism* (PPL), and *anti-pluralism* (MAN), which are likewise latent variables. Figure 10 shows that they are modeled as pairwise correlated latent variables each with three pairwise uncorrelated items. For the items, respondents were again asked to indicate their agreement with a set

Item	Name	Statement
C1	akker6	What people call “compromise” in politics is really just selling out on one’s principles.
C2	cses1	Most politicians do not care about the people.
C3	cses2	Most politicians are trustworthy.
C4	cses3	Politicians are the main problem in [COUNTRY].
C5	cses4	Having a strong leader in government is good for [COUNTRY] even if the leader bends the rules to get things done.
C6	cses5	Most politicians care only about the interests of the rich and powerful.
C7	akker2	The people, and not politicians, should make our most important policy decisions.

Table 6: Statements for items in Hobolt et al.’s (2016) populism model.

of statements which are listed in Table 7. Besides the different content, respondents were also provided with a 7-point scale instead of the 5-point scale used by Hobolt et al. (2016). While the items are modeled as pairwise uncorrelated, Castanho Silva et al. (2018) add an additional technical latent variable (Mtp) to account for the correlation of a subset of items. The loading parameters of this fourth latent variable are constrained to equality. As always, these choices with regard to the structure imply several constraints for the fitting of the model. First, the structure of the loading matrix Λ allows for seven freely estimated loadings with the remaining entries being constrained to 1 for identification purposes or 0 as a result of the model structure. There are six loadings for items relating to the three main latent variables and an additional parameter that determines the loadings with respect to Mtp due to the equality constraint. Second, Φ is freely estimated with the exception of covariances between Mtp and any of the remaining latent variables, which are set to 0. Finally, the off-diagonal entries of Ψ are constrained to zero as a result of the choice of pairwise uncorrelated specific variances. Although this step is not of particular interest for the purpose of this thesis, one might wonder how we can obtain a single measure of populism from this model where none of the latent variables is populism. Castanho Silva et al. (2018) propose aggregate the three main latent variables outside the CFA framework by rescaling them to the interval $[0, 1]$ and then taking their product.

As part of their comparison, Castanho Silva et al. (2020) also assess the “cross-national validity”, i.e. measurement invariance properties, of these and the remaining models. In essence, they test global MI using LR-tests in the MGCFA framework with countries being the relevant groups that are being considered for MI.²⁰ The results of Castanho Silva et al.’s (2018) original analysis are replicated in table 8. In general, the results are not very encour-

²⁰One potential issue is that the study uses convenience samples which exhibit a gender imbalance for some countries. This indicates that the samples may be representative with regard to their respective country population to a different degree across countries. Measurement non-invariance could theoretically also originate from these differences and show up when distinguishing between countries as a result of these imbalances. However, the purpose of my contribution to this paper is less the empirical validity of the overall comparison, but rather whether the detection methods can be fruitfully used to detect non-invariant items in CFA models using real-world data.

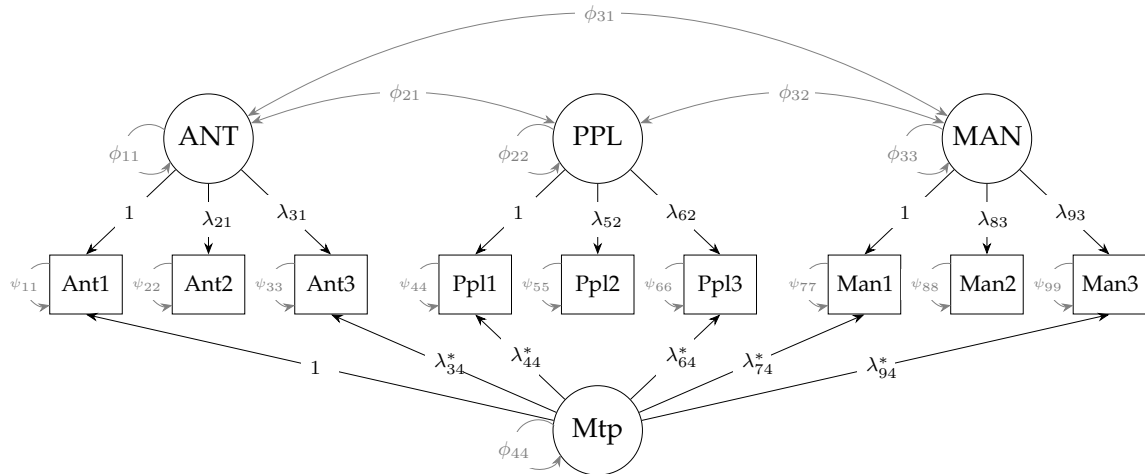


Figure 10: Graphical depiction of Castanho Silva et al.'s (2018) populism model. Note that intercepts were omitted in this visualization. Further note that the five parameters relating to Mtp with an asterisk are constrained to equality.

Item	Name	Statement
Ant1	antiel23	The government is pretty much run by a few big interests looking out for themselves.
Ant2	rwpop8	Government officials use their power to try to improve people's lives.
Ant3	antiel21	Quite a few of the people running the government are crooked.
Ppl1	gewill17	Politicians should always listen closely to the problems of the people.
Ppl2	simple8	Politicians don't have to spend time among ordinary people to do a good job.
Ppl3	gewill3	The will of the people should be the highest principle in this country's politics.
Man1	manich15	You can tell if a person is good or bad if you know their politics.
Man2	manich13	The people I disagree with politically are not evil.
Man3	manich14	The people I disagree with politically are just misinformed.

Table 7: Statements for items in Castanho Silva et al.'s (2018) populism model.

aging: The null hypothesis of metric MI is rejected for all but two models while the null hypothesis of scalar MI (results not included) is rejected for all models. In other words, all models exhibit clear violations of MI, rendering cross-country comparisons meaningless. Additionally, Castanho Silva et al. (2020) stress that one of the two models satisfying metric MI, Elchardus and Spruyt's (2016) model, has an overall poor fit such that the fact that it is measurement invariant with regard to the loadings merely indicates that it fits equally poorly across countries. Thus, they come to the conclusion that only Castanho Silva et al.'s (2018) model has at least some meaningful cross-national validity.

With these clear violations of MI across the board, there may be considerable use in trying to detect the items that contribute most to global MI. As was mentioned before, the used survey items differ by measurement model. Generally, new proposed models have

Model	χ^2 base	χ^2 metric	Difference (deg.free)	p-value
Akkerman et al. (2014)	230.76	297.15	59.935 (40)	.022
Hobolt et al. (2016)	462.03	570.89	89.458 (48)	< .001
Oliver and Rahn (2016)	942.62	1176.99	187.29 (72)	< .001
Elchardus and Spruyt (2016)	229.16	254.76	20.845 (24)	.648
Schulz et al. (2018)	360.48	496.06	116.64 (64)	< .001
Stanley (2011)	629.81	809.27	131.05 (64)	< .001
Castanho Silva et al. (2018)	440.58	599.12	102.04 (88)	.145

Table 8: Tests of global metric MI of the populism models. Source: Castanho Silva et al. (2020)

introduced new survey questions to create these items; only in a few cases have existing survey questions been recycled for new populism models (Castanho Silva et al., 2020). While these items as well as the measurement model is generally justified by theory, most models were at least in part developed empirically by deleting items with small loadings (Castanho Silva et al., 2020). It appears that considerations with regard to the MI properties of potential items are not present in the scale development despite the fact that many of these studies were ultimately interested in cross-country comparisons. This is exactly where the detection methods in this thesis can be applied: Instead of just selecting items that load strongly on their latent variables, researchers that are interested in cross-group comparisons can take into account the measurement invariance properties of each item. Applying the detection methods to their models gives them an idea of which items contribute to subsequent global MI violations. Ideally, this is done at an early stage of the research, i.e. particularly before large-scale surveys are fielded, such that survey questions can be replaced or altered. In that regard, the following application departs from reality: All it can provide is an illustration of what the initial results for applied researchers using these detection methods may look like and how they can be interpreted. Further development of the items then requires substantive field knowledge, as it always is the case with CFA.

6.2 Implementation of Detection Methods

In the previous sections, the detection methods were introduced and used for single-factor models for the sake of clarity. However, several of the populism models, including Castanho Silva et al.'s (2018) model are multidimensional. Fortunately, generalizing the detection methods to the multi-factor case is straightforward: In most cases, the detection methods can simply be applied individually to each latent variable and the items it relates to while leaving the other latent variables and items in the model untouched. For example, for the CR approach, the triangle heuristic is simply applied for each latent variable by systematically changing the marker and reference item in the subset of items that relate to that latent variable. As a result, the methods need to be applied k times in most cases. Generalization is even more straightforward for the BV approach where still only the deletion

of each item is required. However, there are a few discretionary choices that need to be mentioned: First, if an item is modeled to relate to more than one latent variable, it is classified as being non-invariant if it is found to be non-invariant for any of its relationships with a latent variable. Second, some of the models required for the internal model comparison of some of the methods may not be identified. As a default, all items are assumed to be invariant such that in this case items for which the unidentified comparison model was generated is arbitrarily classified as being invariant. Similarly, in the stepwise method R2, an item isn't removed if it is the last one remaining for a given latent variable. Third, for the R approaches, the Bonferroni correction also takes into account the additional tests from all latent variables. Finally, all implementations disregard latent variable-item relationships for which loadings were constrained to a constant or to equality with another parameter. For example, in Castanho Silva et al.'s (2018) model, all loadings that relate to the latent variable Mtp are constrained to equality. Therefore, the latent variable is disregarded by the detection methods. However, note that the items relating to Mtp may still be classified as being non-invariant by analyzing their relationships with the remaining latent variables. The reason why these relationships are disregarded is that such modeling decisions require lots of substantive knowledge and should only be imposed if there are very good reasons to do so. In the model development process, these decisions should therefore come after considerations with regard to MI.

Moreover, for the MInd, CR, and BV approaches, I implement additional metric MI versions of the methods. The motivation for these versions is the fact that Castanho Silva et al. (2020) find that scalar MI isn't achieved by any of the models. Recall that in the previous sections, violations of scalar and metric MI at the item level were lumped together because researchers generally don't know a priori which type of MI may be violated at the item level. In a nutshell these metric MI versions only consider the MI properties of a given item with respect to its loadings. For example, recall that in the standard implementation of the MInd approach, the baseline model constrains the loadings and intercepts across groups to equality and computes a modification index for simultaneously lifting both types of constraints for each item. The metric version now only constrains the loadings for the baseline model and modification indexes refer to the lifting of single loading constraints. All other steps of the methods remain unchanged compared to their original version. In general, the expectation is that Unfortunately the R approaches cannot distinguish between metric and scalar non-invariance which is one of their few disadvantages.

6.3 Results

For the most part, the results with regard to the MI properties of the various models aren't very encouraging. However, this was to be expected given that globally, no model achieves scalar MI and only two achieve metric MI (Castanho Silva et al., 2020). In the following, I focus on the models by Hobolt et al. (2016) and Castanho Silva et al. (2018) which can be viewed as examples of simple and complex models, respectively, with the former violat-

ing scalar and metric and the latter only scalar MI. The results for the remaining models are included in section 8.4 of the appendix. Tables 9 and 10 contain the results for the two models. The first five columns indicate the items that were classified as being non-invariant with the detection methods set to detect metric and scalar non-invariance. For the final three columns, the metric MI versions were used where only the invariance properties with respect to the loadings of each item are studied.

Item	Survey	Metric & Scalar					Metric		
		MInd	CR	BV	R1	R2	MInd	CR	BV
C1	akker6	•		•	•	•	•	•	•
C2	cses1		•						
C3	cses2.r	•	•	•	•	•	•		•
C4	cses3	•	•		•		•		
C5	cses4	•	•	•	•	•	•	•	•
C6	cses5		•					•	
C7	akker2	•	•	•	•	•		•	

Table 9: Items classified as non-invariant (•) in Hobolt et al. (2016).

Table 9 contains the results for the model by Hobolt et al. (2016). At first glance, there is considerable disagreement between the different detection methods: No item is unanimously classified as either invariant or non-invariant by all detection methods. Nonetheless, some items appear to be less problematic than others. In terms of metric and scalar MI, C2 and C6 are only classified as non-invariant by the CR approach which had very low specificity in the simulation study. Interestingly, the questions in the survey that correspond to these two items both concern respondents' beliefs about politicians' interests. Considering the BV and R2 approach, which performed best in the simulation study, the results show that they fully agree in their classifications which is encouraging. According to these two methods, C4 can also be considered to be invariant. Turning to the final three columns, things look slightly better when considering metric MI alone: only C1, C3, and C5 seem really problematic with two or all of the detection methods detecting them as being non-invariant and the most trusted detection method, the BV approach, only picking these three. From a model developing perspective, the results thus indicate that these three items should be studied in more detail and if possible be replaced, modified, or removed.

Similarly, Table 10 contains the results for Castanho Silva et al.'s (2018) model. Again, there is some considerable variation across the methods. Starting with metric and scalar invariance, every item is classified as non-invariant by a majority of the detection methods with the exception of item Man1 which is only selected by the MInd and R1 approach. Recall that these two methods exhibited very low specificity in the simulation study. It thus seems plausible that Man1 is indeed non-invariant. Unfortunately, contrary to the results for the Hobolt et al.'s (2016) model, the best performing detection methods BV and R2 disagree on almost every item. Therefore, it seems prudent to refrain from further analysis of the metric and scalar invariance of the remaining items. Turning to metric MI, things again look slightly more promising: even the MInd approach with its low specificity

Item	Survey	Metric & Scalar					Metric		
		MInd	CR	BV	R1	R2	MInd	CR	BV
Ant1	antiel23	•			•	•			
Ant2	rwpop8.r	•	•	•	•		•	•	•
Ant3	antiel21	•	•		•	•	•		
Ppl1	gewill17	•			•	•	•		
Ppl2	simple8.r	•	•	•	•		•	•	•
Ppl3	gewill3	•	•		•	•	•	•	
Man1	manich15	•			•				
Man2	manich13.r	•	•	•	•	•			•
Man3	manich14	•	•		•	•	•	•	

Table 10: Items classified as non-invariant (•) in Castanho Silva et al. (2018).

classifies several items as being invariant. Most clearly, Ant1 and Man1 are likely invariant with regard to their corresponding loadings. Focusing on the most trustworthy detection method, the BV approach, the most problematic items appear to be Ant2, Ppl2, and Man2. These should be the starting point for further model development. It is noteworthy that they items are classified as non-invariant even though the global null hypothesis of metric MI wasn't rejected. This goes to show that even in such cases, the detection methods may be beneficial and provide further insight by a more fine-grained look at the items' MI properties.

6.4 Summary and Discussion

The results have shown that at least in principle, the detection methods can be applied to more complex CFA models than those used in the simulation study. For both models, the detection methods identified a selection of items that seem particularly problematic in terms of their cross-national validity. However, substantial disagreement in classifying the same items exists between the different detection methods. The fact that the methods perform differently well is part of an explanation, but even the two best methods, BV and R2, can disagree massively. This casts some doubt on how well the methods actually perform in the field. However, given the abysmal global cross-national validity of the populism models, the application of the detection methods is only a starting point of model development in terms of MI. It is therefore not too surprising that there is disagreement between the different detection methods or that they classify many items as being non-invariant. In short, the results should be taken with a grain of salt. Notwithstanding, they are not completely useless: There are some items for which almost all detection methods agree on their being invariant. In substance, researchers can try to generate plausible explanation as to why they may be different from the remaining items and improve their models by using these insights for devising new replacement items for those items that are clearly non-invariant. In the context of Castanho Silva et al.'s (2020) study, another option and step forward would be to draw on the remaining items that were obtained for the remaining

populism models as replacements.

7 Conclusion

7.1 Recap

7.2 Open questions

- when should researchers start looking at item level? Only when global H_0 rejected?

7.3 Limitations/further research

- detection method ranking items by their degree of MI. already possible for BV, MInd, R
- simulation assumes completely independent indicators, could be extended =; more general: more complex CFA models, even if still single-factor, but also multi-factor models.

References

- Akkerman, A., Mudde, C., & Zaslove, A. (2014). How populist are the people? measuring populist attitudes in voters. *Comparative Political Studies*, (9), 1324–1353.
- 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.
- Beauducel, A. (2007). In spite of indeterminacy many common factor score estimates yield an identical reproduced covariance matrix. *Psychometrika*, 72(3), 437–441.
- 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.
- Castanho Silva, B., Andreadis, I., Anduiza, E., Blanuša, N., Corti, Y. M., Delfino, G., Rico, G., Ruth, S. P., Spruyt, B., Steenbergen, M., & Littvay, L. (2018). Public opinion surveys: A new scale. In K. Hawkins, R. Carlin, L. Littvay, & C. R. Kaltwasser (Eds.), *The ideational approach to populism: Concept, theory, and analysis*. Routledge.
- Castanho Silva, B., Jungkunz, S., Helbling, M., & Littvay, L. (2020). An empirical comparison of seven populist attitudes scales. *Political Research Quarterly*, 73(2), 409–424. <https://doi.org/10.1177/1065912919833176>
- 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
- 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
- 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>
- Elchardus, M., & Spruyt, B. (2016). Populism, persistent republicanism and declinism: An empirical analysis of populism as a thin ideology. *Government & Opposition*, (1), 111–133.
- Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, methods and applications*. Springer.
- Hershberger, S. L. (2014). Factor score estimation. *Wiley statsref: Statistics reference online*. John Wiley & Sons, Ltd. <https://doi.org/https://doi.org/10.1002/9781118445112.stat06532>
- Hobolt, S., Anduiza, E., Carkoglu, A., Lutz, G., & Sauger, N. (2016). *Democracy divided? people, politicians and the politics of populism*. http://www.cses.org/plancom/module5/CSES5_ContentSubcommittee%20FinalReport.pdf
- 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.
- Johnson, D. R., & Creech, J. C. (1983). Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 48(3), 398–407. <http://www.jstor.org/stable/2095231>
- 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
- 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
- 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.
- Mudde, C. (2004). The populist zeitgeist. *Government and opposition*, 39(4), 541–563.
- 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
- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, 38(2), 171–189. <https://doi.org/10.1111/j.2044-8317.1985.tb00832.x>
- Muthén, B., & Asparouhov, T. (2013). Bsem measurement invariance analysis. *Mplus Web Notes*, 17 (Jan 11). <http://www.statmodel.com/examples/webnotes/webnote17.pdf>
- Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. W.H. Freeman; Co.
- Oliver, J. E., & Rahn, W. M. (2016). Rise of the trumpenvolk: Populism in the 2016 election. *Annals of the American Academic of Political and Social Science*, (1), 189–206.
- 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>
- R Core Team. (2020). R: A language and environment for statistical computing [version 4.0.2]. *R Foundation for Statistical Computing*. <https://www.R-project.org/>
- 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/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>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling [v0.6-9]. *Journal of Statistical Software*, 48(2), 1–36. <https://www.jstatsoft.org/v48/i02/>

- 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.
- Schulz, A., Müller, P., Schemer, C., Wirz, D. S., Wettstein, M., & Wirth, W. (2018). Measuring populist attitudes on three dimensions. *International Journal of Public Opinion Research*, (2), 316–326.
- Sörbom, D. (1989). Model modification. *Psychometrika*, 54(3), 371–384. <https://doi.org/10.1007/BF02294623>
- Stanley, B. (2011). Populism, nationalism, or national populism? an analysis of slovak voting behaviour at the 2010 parliamentary election. *Communist and Post-Communist Studies*, (4), 257–270.
- 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.*
- Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models.
- Thomson, G. H. (n.d.). *The factorial analysis of human ability*. University of London Press.
- 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

W.l.o.g., assume an EFA model for centered items \mathbf{Y}

$$\mathbf{Y} = \mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \quad (63)$$

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

$$\boldsymbol{\Sigma}(\theta) = \mathbf{\Lambda}\boldsymbol{\Phi}\mathbf{\Lambda}^\top + \boldsymbol{\Psi}, \quad (64)$$

with $\theta := (\mathbf{\Lambda}, \boldsymbol{\Psi})$ because $\boldsymbol{\Phi}$ is an identity matrix.

Given a sample of n observations and p items, we can write the likelihood for single observations $i = 1, \dots, n$ as

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

and the joint likelihood for n observations as

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

Thus, the joint log-likelihood is given by

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

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

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

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 \mathbf{S} as just \mathbf{S} because it's within the trace. As a result, the log-likelihood can be rewritten as a function of the model-implied covariance matrix and the sample covariance matrix which is a sufficient statistic with respect to the parameters of the EFA model:

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

where the addition of the constants $\log \det(\mathbf{S})$ and p conveniently sets the log-likelihood to zero when $\Sigma = \mathbf{S}$, resulting in the fit function $F(\cdot)$.

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 Σ . Suppose we have a positive definite matrix $\mathbf{V} \sim \mathcal{W}_p(\Sigma, n)$ with pdf

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

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

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

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

8.2 Comparative Fit Index (CFI)

The comparative fit index (CFI) (Bentler, 1990) is a measure of model fit for a given model \mathcal{M} . Let $\mathcal{M}_{\text{base}}$ be the corresponding baseline model for which all items are modelled to be independent without any underlying latent variables. In other words, $\mathcal{M}_{\text{base}}$ is a model with only item intercepts and variances. Formally, the CFI can be defined as

$$\text{CFI}(\mathcal{M}) := 1 - \frac{\max\{T(\mathcal{M}), 0\}}{\max\{T(\mathcal{M}), T(\mathcal{M}_{\text{base}}), 0\}} \in [0, 1] \quad (65)$$

where $T(\mathcal{M})$ is defined as

$$T(\cdot) := nF(\mathcal{M}) - \text{df}_{\mathcal{M}} \quad (66)$$

and analogously for $\mathcal{M}_{\text{base}}$. Since $nF(\cdot)$ is the χ^2 -statistic defined in equation (19) and because of the analogy to the likelihood, the CFI can be interpreted as an adjusted likelihood ratio of these two models on the interval $[0, 1]$.

8.3 Scale invariance of the EFA model

8.4 Non-invariance Classifications for Remaining Models in Castanho Silva et al. (2020)

[Convergence issue for some sub-models in CR for basemodels of OR & Stan: unclear whether result of base models or lavaan.]

Survey	MInd	CR	BV	R1	R2
akker1	•			•	•
akker2	•	•		•	
akker3		•			
akker4	•	•	•	•	•
akker5	•	•	•	•	•
akker6	•	•	•	•	•

Table A1: Items classified as non-invariant (•) in Akkerman et al. (2014).

Survey	MInd	CR	BV	R1	R2
es1	•		•		
es2	•	•	•	•	•
es3	•	•			•
es4	•	•	•	•	•

Table A2: Items classified as non-invariant (•) in Elchardus and Spruyt (2016).

Survey	MInd	CR	BV	R1	R2
ow_ae1	•	-		•	
ow_ae2	•	-		•	•
ow_ae3	•	-		•	•
ow_ae4	•	-		•	•
ow_ae5	•	-	•	•	•
ow_me1	•	-		•	•
ow_me2	•	-		•	
ow_me3	•	-	•	•	•
ow_me4	•	-	•	•	•
ow_na1	•	-		•	
ow_na2	•	-	•	•	•
ow_na3	•	-	•	•	•

Table A3: Items classified as non-invariant (•) in Oliver and Rahn (2016). Note that CR results are excluded due to convergence issues.

Survey	MInd	CR	BV	R1	R2
nccr_ant1	•			•	•
nccr_ant2		•			
nccr_ant3	•	•		•	•
nccr_sov1	•				
nccr_sov2	•	•			
akker2	•	•		•	•
nccr_hom1	•		•	•	•
nccr_hom2	•	•	•	•	
nccr_hom3	•	•	•	•	•

Table A4: Items classified as non-invariant (•) in Schulz et al. (2018).

Survey	MInd	CR	BV	R1	R2
stanley1	•	-		•	
stanley2	•	-			
stanley3	•	-		•	•
stanley4	•	-	•	•	•
stanley5	•	-		•	•
stanley6	•	-	•	•	•
stanley7	•	-	•	•	•
stanley8	•	-		•	

Table A5: Items classified as non-invariant (•) in Stanley (2011). Note that CR results are excluded due to convergence issues.



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.