Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Winter 2021/2022

Pit Rieger

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

February 10, 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, two versions of a novel detection approach are introduced. 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. Second, the performance of these two methods and four existing methods is assessed in a simulation study. This enables a comparison of detection methods and offers guidance for choosing a method in applied research. Finally, the detection methods are applied to different 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 the existing methods have difficulty detecting non-invariant items reliably. Of the four existing detection methods, only one can be recommended conditionally. At the same time, one version of the novel approach performs very well across all settings and can thus be recommended more generally. 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://prieger.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/pitrieger/masterthesis.

Contents

List of Abbreviations

ANOVA Analysis of variance.

BV Byrne & Van de Vijver.

CFA Confirmatory factor analysis.

CFI Comparative fit index.

CR Cheung & Rensvold.

DGP Data-generating process.

DIF Differential item functioning.

DTF Differential test functioning.

EFA Exploratory factor analysis.

FN False negative.

FP False positive.

IRT Item response theory.

J Janssens.

LR Likelihood ratio.

MGCFA Multi-group confirmatory factor analysis.

MI Measurement invariance.

MInd Modification index.

ML Maximum likelihood.

MLE Maximum likelihood estimate.

R Rieger.

RMSEA Root mean square error of approximation.

TN True negative.

TP True positive.

1 Introduction

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, 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 processed similarly by all groups. 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, even when controlling for the 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 is situated in the second category and specifically focuses on the detection of items that violate MI. These tests are particularly useful in the setting of partial MI (Byrne et al., 1989), i.e. where MI holds for a subset of groups or items, because they can point to the items for 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 between CFA and 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 three-

fold. 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). A simple detection method may contribute to a popularization of the item-wise analysis of MI properties and thereby aid the amelioration of this neglect. Second, the novel approach can easily be visualized. Again, this is helpful for promoting the use of this detection method among applied researchers. Finally, compared to goodness-of-fit-based tests, fewer 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 fitted 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 of 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 detection 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, can even be considered a viable option in some settings, but that it is easily outperformed by one version of the novel approach, the R2 method. While the R2 method is slightly less specific than the BV method, it achieves noticeably higher sensitivity and most importantly works well for all types of MI violation. The remaining three existing methods and the other version of the novel approach perform rather poorly, often due to very low specificity. In light of the neglect of MI issues in empirical research, it is very encouraging to see that the detection methods that perform best are also the most straightforward ones in terms of theory and implementation.

In the application of the detection methods to the CFA models measuring populism, the results show that there can be considerable disagreement among 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, several items that are likely major contributors to the violation of global MI of the models were 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 generalize 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 ($\hat{}$). Nested indices are written in parentheses, e.g. if index j is nested in l, it is written as $l(j)$. 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 j is generally used for latent variables $j = 1, ..., k$, but sometimes for observations $j = 1, ..., n$. These cases are always clearly documented. Moreover, one potentially confusing aspect of the notation in this thesis is that the group index (usually l) is written in superscript. This decision was made in line with the literature (e.g. Davidov et al., 2014) and highlights the focus on the grouping structure. Nonetheless, caution needs to be exercised not to confuse the index with a power.

CFA/EFA model

$\mathcal{M}$	Specific CFA model (structure)
$\boldsymbol{Y}$	Items / indicators
$\boldsymbol{\eta}$	Latent variables
$\boldsymbol{\tau}$	Intercepts
$\boldsymbol{\Lambda}$	Loading matrix
λ_{ij}	Loading
$\boldsymbol{\varepsilon}$	Errors / specific variables
μ	Expectation of latent variable
ϕ	Variance of Latent variable
$\boldsymbol{\Psi}$	Covariance matrix of errors
$\boldsymbol{\Phi}$	Covariance matrix of latent variables
$\boldsymbol{\Sigma}$	Covariance matrix of items
$\boldsymbol{S}$	Sample covariance matrix of items
p	Number of items
k	Number of latent variables
g	Number of groups
i	Item index
j	Latent variable index
l	Group index
n	Number of observations per group
N	Total number of observations
d	Degrees of freedom
S	Set of non-invariant items

Model Estimation

$F(\cdot)$	Fit function
$\mathcal{L}(\cdot)$	Likelihood function
$\ell(\cdot)$	Log-likelihood function
$\Gamma(\cdot)$	Likelihood ratio statistic

Regression of Y on η

γ	Intercept
β	Slope
r	Residuals
ν	Expectation of residuals

Regression of r on η

κ	Intercept
ω	Slope

Simulation

m	Number of non-invariant items
h	Share of groups affected by non-invariance
δ_τ	Magnitude of bias on intercepts

δ_λ	Magnitude of bias on loadings
Distributions	
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate-normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
χ^2_d	Chi-squared distribution with d degrees of freedom
t_d	Student's t-distribution with d degrees of freedom
F_{d_1,d_2}	F-distribution with d_1 and d_2 degrees of freedom
Miscellaneous	
$\mathbb{1}_{\{\cdot\}}$	Indicator function
$\boldsymbol{I}_{p\times p}$	Identity matrix
$\mathbf{0}$	Zero vector/matrix
$\boldsymbol{R}$	Rotation matrix
$\boldsymbol{P}$	Permutation matrix
α	Significance level
$\wp$	p-value

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 circumvents 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 why measurement models are necessary 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 alternative 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 structure, or 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. This should be kept in mind when talking about CFA models.

2.1 Refresher: Exploratory Factor Analysis

2.1.1 Setup

EFA supposes a set of p continuous[1] items $\boldsymbol{Y} = (Y_1, ..., Y_p)^{\mathrm{T}} \in \mathbb{R}^p$, also referred to as *manifest variables* or *indicators*. In the context of survey research, each item relates to a question in a battery of p survey questions that respondents provide answers to. Furthermore, the items are assumed to relate to k latent variables $\boldsymbol{\eta} = (\eta_1, ..., \eta_k)^{\mathrm{T}} \in \mathbb{R}^k$, also referred to as *factors*, following some distribution with $\mathbb{E}[\boldsymbol{\eta}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{\eta}) = \boldsymbol{\Phi}$. For now, the number of latent variables k is assumed to be known. However, in the practical application of EFA, this is typically not the case and how k can be chosen will be discussed in section 2.1.5 below. Further, the relationship is assumed to be of the following linear, multivariate, and multiple regression form:

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

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_p)^{\mathrm{T}}$ are intercepts, λ_{ij} are regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_p)^{\mathrm{T}}$ are errors. In factor analysis, it is more common to refer to the coefficients as *(factor) loadings* and the errors as *specific variables*. Note that the assumed relationship implies that each

[1] Strictly 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).

item is a linear combination of all latent variables plus an intercept and an idiosyncratic error.

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

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

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, additional assumptions are necessary. Particularly, the specific variables are assumed to have mean zero, to be pairwise uncorrelated, and to be uncorrelated with the latent variables:

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \tag{3a}$$
$$\operatorname{Cov}(\boldsymbol{\varepsilon}) = \mathbf{\Psi}, \text{ a diagonal matrix} \tag{3b}$$
$$\operatorname{Cov}(\boldsymbol{\eta}, \boldsymbol{\varepsilon}) = \mathbf{0}, \tag{3c}$$

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

$$\begin{aligned} \mathbf{\Sigma} \coloneqq \operatorname{Cov}(\boldsymbol{Y}) &= \operatorname{Cov}(\mathbf{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}) \\ &\overset{(3c)}{=} \operatorname{Cov}(\mathbf{\Lambda}\boldsymbol{\eta}) + \operatorname{Cov}(\boldsymbol{\varepsilon}) \\ &= \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}^{\mathsf{T}} + \mathbf{\Psi}. \end{aligned} \tag{4}$$

It is clear from this decomposition that the assumption of a relationship of the form shown in equation (2) implies a covariance structure for the items because $\mathbf{\Sigma}$ clearly depends on the model parameters. Thus, it often makes sense to talk about model-implied covariances. I occasionally write $\mathbf{\Sigma}(\mathbf{\Lambda}, \mathbf{\Phi}, \mathbf{\Psi})$ to emphasize this dependence. This is crucial for the estimation of factor analysis models: A good model yields a model-implied covariance matrix that resembles the observed sample covariance matrix of the items. As shown further below, 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 an invertible matrix $\boldsymbol{R}$ of dimension $k \times k$ and let

$$\widetilde{\boldsymbol{\Lambda}} \coloneqq \boldsymbol{\Lambda}\boldsymbol{R} \tag{5a}$$

$$\widetilde{\boldsymbol{\eta}} \coloneqq \boldsymbol{R}^{-1}\boldsymbol{\eta} \tag{5b}$$

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

$$\boldsymbol{Y} = \boldsymbol{\tau} + \widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\eta}} + \boldsymbol{\varepsilon} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{\eta} + \boldsymbol{\varepsilon} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{6}$$

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 $\boldsymbol{\Lambda}$ and $\boldsymbol{\eta}$. Although $\boldsymbol{R}$ 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 solution is to constrain the means and covariances of the latent variables to render EFA identifiable. More specifically, the latent variables are required to have mean zero, unit variance and have to be pairwise uncorrelated, i.e.

$$\mathbb{E}\left[\boldsymbol{\eta}\right] = \boldsymbol{0} \tag{7a}$$

$$\operatorname{Cov}\left(\boldsymbol{\eta}\right) = \boldsymbol{I}_{k\times k}\text{, an identity matrix.} \tag{7b}$$

Given these properties of $\boldsymbol{\eta}$, it is easy to see that $\operatorname{Cov}\left(\widetilde{\boldsymbol{\eta}}\right) = \boldsymbol{I}_{k\times k}$ if and only if $\boldsymbol{R}$ is the identity matrix itself. The resulting unique solution under these constraints is therefore commonly referred to as the *unrotated solution*. This solution can then still be subjected to post-estimation transformations $\boldsymbol{R}$ while yielding model parameters that are equally valid because they only violate the arbitrary constraints on the latent variables.[2] If $\boldsymbol{R}$ is an orthogonal matrix, the transformation preserves the uncorrelatedness of the factors in the unrotated solution and is referred to as an *orthogonal rotation*. Other transformations that are not orthogonal are called *oblique rotations*. Oftentimes, the goal of applying a rotation is to ease the interpretation of the factor loadings. In this regard, different algorithmic rotations have been proposed, which result in loading matrices that are more easily in-

[2]Note that the constraints are arbitrary because the true location and scale of the latent variables cannot be known due to their unobservability. Instead, they are arbitrarily fixed for technical purposes.

terpretable. 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 (for an overview, see Browne, 2001). 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) approach[3] which requires an additional distributional assumption. The usual assumption of a multivariate normal distribution can be written as

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

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

$$\widetilde{\boldsymbol{Y}} \sim \mathcal{N}_p(\boldsymbol{\Lambda\eta}, \boldsymbol{\Sigma}), \tag{9}$$

where in practice, the corresponding sample mean is simply subtracted 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 $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$. It is easy to overlook how astonishing it is that estimates for these quantities can be obtained when considering that $\boldsymbol{\eta}$ is unknown. This is due to the assumptions about $\boldsymbol{\eta}$ which render the likelihood dependent on nothing but $\boldsymbol{\Sigma}$ which in turn depends on the quantities of interest $\boldsymbol{\Lambda}$ and $\boldsymbol{\Psi}$.

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

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

where $\boldsymbol{S}$ is the sample covariance matrix of $\boldsymbol{Y}$. It turns out that $\boldsymbol{S}$ is a sufficient statistic for the parameters in the centered 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-

[3] Another common alternative is the principal factor method. However, of the two, only ML estimation can be used for the estimation of CFA models. I therefore only discuss the ML approach.

ing the fit function

$$F\big(\theta \mid \boldsymbol{S}\big) = \log\det\big(\boldsymbol{\Sigma}(\theta)\big) - \log\det(\boldsymbol{S}) + \operatorname{tr}\big(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\theta)\big) - p \propto -\ell\big(\boldsymbol{\Sigma}(\theta) \mid \boldsymbol{S}\big), \tag{11}$$

where the replacement of constants in equation (10) with $\log\det(\boldsymbol{S})$ and p conveniently sets the fit function to zero when $\boldsymbol{\Sigma}(\theta) = \boldsymbol{S}$, 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 items, are

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

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

2.2 Confirmatory Factor Analysis

2.2.1 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. All prior notation can be kept and the model can still be written in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{13}$$

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

$$\boldsymbol{\Sigma}(\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\mathsf{T}} + \boldsymbol{\Psi}. \tag{14}$$

The consequential difference lies in how the parameters in CFA are treated: To formulate a CFA model, researchers must define the *structure* of their model. By structure, I 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, one may want to formulate a model structure where items load only on a subset of latent variables. Recall that for EFA every item was allowed (or forced) 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 ($\mathbf{\Lambda}$), which latent variables are dependent on each other ($\mathbf{\Phi}$), and which specific variables are dependent on each other ($\mathbf{\Psi}$). 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 $\mathbf{\Phi}$ 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 with 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, $\mathbf{\Phi}$ was further assumed to be 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, one 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, one may want to ensure that two loadings have the same magnitude. Perhaps, this raises the more general question of why one 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 (DGP) 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

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

[5] An 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 in addition a constant loading is imposed on the relationship between said latent variable and another item.

testing in the CFA framework will be 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 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.

To further illustrate the implications of model structures and the constraints they imply, I 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-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, 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 issue 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 $p = 6$ items. 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 while support of the left tends to be indicative of more liberal and green positions. However, the proposed structure for the factor loadings indicate that the researchers believe that their items 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 such that the errors can be modeled as pairwise uncorrelated. Denoting these considerations as

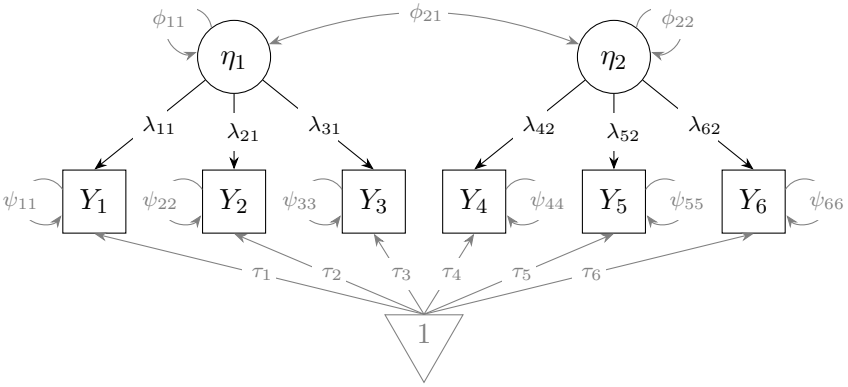

Figure 1: *CFA model with two correlated latent variables and six manifest variables.*

model structure $\mathcal{M}$, they can be formally reflected in $\boldsymbol{\tau}_{\mathcal{M}}$, $\boldsymbol{\Lambda}_{\mathcal{M}}$, $\boldsymbol{\Phi}_{\mathcal{M}}$, and $\boldsymbol{\Psi}_{\mathcal{M}}$ as follows:

$$\boldsymbol{\tau}_{\mathcal{M}} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix}, \boldsymbol{\Lambda}_{\mathcal{M}} = \begin{bmatrix} \lambda_{11} & \\ \lambda_{21} & \\ \lambda_{31} & \\ & \lambda_{42} \\ & \lambda_{52} \\ & \lambda_{62} \end{bmatrix}, \boldsymbol{\Phi}_{\mathcal{M}} = \begin{bmatrix} \phi_{11} & \phi_{21} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \boldsymbol{\Psi}_{\mathcal{M}} = \begin{bmatrix} \psi_{11} & & \\ & \ddots & \\ & & \psi_{66} \end{bmatrix}.$$

In this structure, blank matrix entries reflect zero constraints. To emphasize, the structure implies that the intercepts are all freely estimated. Further, $\boldsymbol{\Lambda}_{\mathcal{M}}$ 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 $\boldsymbol{\Phi}_{\mathcal{M}}$ are the result of allowing the latent variables to be correlated. Finally, $\boldsymbol{\Psi}_{\mathcal{M}}$ is a diagonal matrix because of the assumption that all dependencies of the items can be explained by their respective relationships with the latent variables. This also 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 conventions regarding the shapes for latent and manifest variables. Specifically, latent variables are represented with circles, manifest items with squares and intercepts with a triangle. Furthermore, straight single-headed arrows represent a linear relationship with a given path coefficient, circular arrows represent variances, and the double-headed arrow between η_1 and η_2 represents their covariance. The 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 CFA models is done with a ML approach. Given that the difference between CFA and EFA can be reduced to parameter

constraints, all that changes is that these constraints are taken into account in the maximization of the likelihood: For example, parameters that are constrained to zero are held fixed at zero for the ML estimation. Similarly, a single parameter would be optimized 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 dependes on k and is achieved with the help of an unrotated solution that confines $\mathbf{\Phi}$ 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 $\mathbf{\Lambda}$ will ensure in most cases that there is no issue with rotational invariance. What remains is the issue of scaling of the latent variables for which several alternatives exist (for an overview, see Little et al., 2006). 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 item (*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 there are p items in a model structure which is denoted as $\mathcal{M}$. 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,

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

which is independent of any other aspects than p in $\mathcal{M}$.

On the other hand, the number of free parameters d_{free} is given by the sum of the number

of independent and non-constant parameters for intercepts, loadings, factor covariances, and unique variances in $\mathcal{M}$. In total, any model structure contains $2p + pk + k(k+1)/2$ parameters that are either constrained or free. The constrained parameters may be constrained as a result of modeling choices such as equating a set of parameters or fixing them to a constant. However, constraints also include those that are necessary for model identification. Let $d_{\text{constrained}}$ denote the number of fixed parameters, then

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

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

$$d_{\mathcal{M}} = d_{\text{known}} - d_{\text{free}}. \tag{17}$$

Counting these quantities by hand can quickly get tedious for more complex models. While they are provided by implementations of CFA such as `lavaan` by default, it can still be instructive to count them by hand.

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, it can be used to 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 items. 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 $\theta_{\mathcal{M}} \coloneqq (\boldsymbol{\Lambda}_{\mathcal{M}}, \boldsymbol{\Phi}_{\mathcal{M}}, \boldsymbol{\Psi}_{\mathcal{M}})$ denote the parameters of said model and $\hat{\theta}_{\mathcal{M}}$ their maximum likelihood estimates (MLE). The global hypothesis of the model-implied covariance matrix being equal to the sample covariance matrix is

$$H_0 : \quad \boldsymbol{\Sigma}\left(\theta_{\mathcal{M}}\right) = \boldsymbol{S}. \tag{18}$$

Jöreskog (1969) gives a statistic for testing this hypothesis in terms of the fitting function $F(\cdot)$ as[6]

$$nF\left(\boldsymbol{\Sigma}\left(\hat{\theta}_{\mathcal{M}}\right) \;\middle|\; \boldsymbol{S}\right) \overset{H_0}{\sim} \chi^2_{d_{\mathcal{M}}}. \tag{19}$$

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

In the second case, the comparison of two nested models, the standard LR test can be used. Suppose a model structure $\mathcal{M}_2$ which is nested in $\mathcal{M}_1$ and the following hypothesis

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

A corresponding LR statistic Γ is given by

$$\Gamma(\mathcal{M}_1, \mathcal{M}_2) \coloneqq 2\Bigg(F\left(\boldsymbol{\Sigma}\left(\hat{\theta}_{\mathcal{M}_2}\right) \;\middle|\; \boldsymbol{S}\right) - F\left(\boldsymbol{\Sigma}\left(\hat{\theta}_{\mathcal{M}_1}\right) \;\middle|\; \boldsymbol{S}\right)\Bigg) \overset{H_0}{\sim} \chi^2_{(d_{\mathcal{M}_1} - d_{\mathcal{M}_2})} \tag{21}$$

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.[7]

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 the purposes of most CFA models, 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, three items `Y1`, `Y2`, and `Y3`, each loading on `eta`, 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) instead of the other way around.

[7] https://lavaan.ugent.be/tutorial/index.html

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

```
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. Removing covariances that are included by default 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
```

which would correspond to the following model in mathematical notation:

$$
\begin{aligned}
Y_1 &= \tau_1 + \lambda_{11}\eta_1 + \varepsilon_1 \\
Y_2 &= \tau_2 + \lambda_{21}\eta_1 + \varepsilon_2 \\
Y_3 &= \tau_3 + 2\eta_1 + \varepsilon_3 \\
Y_4 &= \tau_4 + a\eta_1 + \lambda_{42}\eta_2 + \varepsilon_4 \\
Y_5 &= \tau_5 + \lambda_{52}\eta_2 + \varepsilon_5 \\
Y_6 &= \tau_6 + a\eta_2 + \varepsilon_6.
\end{aligned}
$$

Adding these coefficients in the `lavaan`-notation thus 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*` "combines" the loadings of `eta1` on `Y4` and of `eta2` on `Y6`. Thus, a single parameter `a` 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, the parameters that are to be combined 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 some loading, although it becomes increasingly difficult to justify such structural modeling decisions.

Finally, the third operator, `~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 i

```
Yi ~1
```

or with the use of an argument of the CFA function.[8] 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 desired model structure.

2.2.6 Factor Extraction

Up to this point, this section was mostly concerned with the structure, fitting and testing of 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 the 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 with factor analysis, multiple options exist for the *extraction* of latent variables (for an overview, see Hershberger, 2014). However, for practical purposes, most methods yield comparable results (c.f. Beauducel, 2007). A commonly used method is Thomson's (1939) regression method which is also the `lavaan` default. Given a fitted

[8] `meanstructure = TRUE`

model with estimates $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Phi}}$, and $\hat{\mathbf{\Psi}}$ as well as the covariance matrix $\hat{\mathbf{\Sigma}}$ which they imply, the latent variables $\hat{\boldsymbol{\eta}}$ can be estimated with the regression method via

$$\hat{\boldsymbol{\eta}} = \widetilde{\boldsymbol{Y}}\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Lambda}}\hat{\mathbf{\Phi}} \tag{22}$$

where $\widetilde{\boldsymbol{Y}}$ 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 $n \times k$.

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, I further develop the example of researchers trying to study political ideology. Then, a formal definition of different types of MI is given. Finally, the section is concluded with a discussion of how one can test for violations of global MI, i.e. at the model 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 *is the average Swiss citizen more culturally conservative than the average German*. 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. An item 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, this violation can be visualized. Suppose that the case described above applies to two groups, A and B, and item Y_4. 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 ignoring 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.

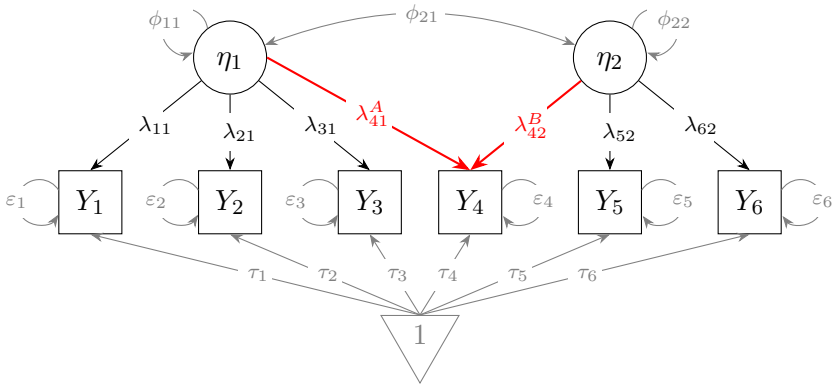

Figure 2: *Violation of configural invariance in the model shown in Figure 1.*

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 loadings 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 different 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 their item responses (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 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, 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 (c.f. Davidov & De Beuckelaer, 2010). 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 $\boldsymbol{Y}$ is independent of the group $l = 1, ..., g$ given $\boldsymbol{\eta}$:

$$\boldsymbol{Y} \perp\!\!\!\perp l \mid \boldsymbol{\eta}. \tag{23}$$

Letting $f(\boldsymbol{Y})$ denote the distribution of $\boldsymbol{Y}$, it is equivalent to write

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

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).

A formal definition of these types first requires the notion of *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 of the same structure for all groups $l = 1, ..., g$. This is equivalent to fitting g independent CFA models of the form

$$\boldsymbol{Y}^l = \boldsymbol{\tau}^l + \Lambda^l \boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l, \tag{25}$$

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, every 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, the parameters in $\boldsymbol{\Lambda}$ that are zero-constrained by the

model structure are required to hold across groups. Formally, let λ_{ij}^{l} denote the i^{th} item's loading on the j^{th} latent variable in group l. Then, given the proposed structure $\mathbf{\Lambda}$ across all groups, configural invariance is satisfied if

$$\mathbb{1}_{\{\lambda_{ij}^{1}=0\}} = \mathbb{1}_{\{\lambda_{ij}^{2}=0\}} = ... = \mathbb{1}_{\{\lambda_{ij}^{g}=0\}} \; \forall \; i = 1, ..., p \; \& \; j = 1, ..., k, \tag{26}$$

where $\mathbb{1}_{\{.\}}$ is an indicator function. By now, it should be clear that the example shown in Figure 2 is a clear violation of configural invariance. Specifically, in the example, configural invariance is violated because

$$\begin{aligned} &\mathbb{1}_{\{\lambda_{41}^{A}=0\}} = 0 \neq 1 = \mathbb{1}_{\{\lambda_{41}^{B}=0\}} \; \& \\ &\mathbb{1}_{\{\lambda_{42}^{A}=0\}} = 1 \neq 0 = \mathbb{1}_{\{\lambda_{42}^{B}=0\}}. \end{aligned} \tag{27}$$

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

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

$$\lambda_{ij}^{1} = \lambda_{ij}^{2} = ... = \lambda_{ij}^{g} \; \forall \; i = 1, ..., p \; \& \; j = 1, ..., k. \tag{28}$$

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 latent variable estimates 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 be constant across groups such that

$$\tau_{i}^{1} = \tau_{i}^{2} = ... = \tau_{i}^{g} \; \forall \; i = 1, ..., p. \tag{29}$$

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 items when all types of MI hold (c.f. Davidov et al., 2014). Yet, for completeness' sake, note that additional types of MI exist. Recall that the item-specific errors $\boldsymbol{\varepsilon}$ follow a p-dimensional distribution with $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Psi}$. Another type of MI, namely *residual invariance* would then require that the covariances hold across groups such that

$$\boldsymbol{\Psi}^1 = \boldsymbol{\Psi}^2 = ... = \boldsymbol{\Psi}^g. \tag{30}$$

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 chosen 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.[9] *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 re-

[9] In the literature on CFA, it appears to be much more common to focus on MI as a model property. While this makes sense because the global MI properties ultimately determine the validity of inference on the basis of the model, it is still a shortcoming of the literature. 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 have prominent names and are 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. However, note that there have been attempts to unify the research on MI, DTF, and DIF (e.g. Stark et al., 2006).

lated: The presence of non-invariant items implies non-invariance at the model-level and vice versa. However, whether a single non-invariant item results in a meaningful degree of global non-invariance is an empirical question that depends on many different aspects of the model and data. Looking at it from the other end, the real benefit of thinking about MI at the item-level lies in the fact that it can give cues where to begin with model improvement if violations of MI were found globally.

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. In other words, they propose that the relevant parameters are estimated freely for each group 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 MI 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 one wants to compare on the grounds of theoretical considerations. Second, these options should not be viewed as mutually exclusive. To the contrary, one should always study or at least consider potential sources of

non-invariance. Moreover, the options can and sometimes have to be combined, for instance by subsetting to groups that exhibit less MI and in a next step removing/replacing a non-invariant 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 Tests 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 equation (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 equation (25). The fully unconstrained model is therefore much more flexible than the fully constrained model.

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, the fully unconstrained model in equation (25) can be taken as a baseline model which is incrementally imposed with 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, the invariance of said family of parameters can be inferred. The underlying intuition is that 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. Vice versa, constraining parameters that violate MI would result in a significant decrease in goodness-of-fit.

In the following, the baseline model for a given model structure $\mathcal{M}$ is referred to as $\mathcal{M}^{\text{base}}$ and has the form shown in equation (25). As families of parameters, the loadings will be considered for metric invariance and both the loadings and intercepts for scalar invariance. For testing global metric invariance, constraining all loadings in $\mathcal{M}^{\text{base}}$ to equality across groups yields the *weakly constrained model*:

$$\mathcal{M}^{\text{weak}}: \quad \boldsymbol{Y}^l = \boldsymbol{\tau}^l + \boldsymbol{\Lambda}\boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l, \tag{31}$$

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*

$$\mathcal{M}^{\text{strong}}: \quad \boldsymbol{Y}^l = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta}^l + \boldsymbol{\varepsilon}^l. \tag{32}$$

Letting $\boldsymbol{\Sigma}(\mathcal{M})$ denote the model-implied covariance of a model with structure $\mathcal{M}$, these three models can be related to hypotheses corresponding to metric and scalar invariance, which should be tested sequentially due to the hierarchy of MI types.

$$\text{Weak MI} \quad H_0^{\text{weak}}: \quad \boldsymbol{\Sigma}\left(\mathcal{M}^{\text{weak}}\right) = \boldsymbol{\Sigma}\left(\mathcal{M}^{\text{base}}\right) \tag{33}$$

$$\text{Strong MI} \quad H_0^{\text{strong}}: \quad \boldsymbol{\Sigma}\left(\mathcal{M}^{\text{strong}}\right) = \boldsymbol{\Sigma}\left(\mathcal{M}^{\text{weak}}\right) \tag{34}$$

Because of the nested nature of the three models, the LR statistic, defined in equation (21), can be used:

$$\Gamma\big(\mathcal{M}^{\text{base}}, \mathcal{M}^{\text{weak}}\big) \quad \overset{H_0^{\text{weak}}}{\sim} \quad \chi^2_{(g-1)d_{freeload}}. \tag{35}$$

where, $d_{freeload}$ is the number of free loading parameters in the model structure such that $(g-1)d_{freeload}$ is the difference in the number of parameters that have to be estimated for the two models.

Analogously, for the comparison of the strongly-constrained with the weakly-constrained model, the following statistic applies:

$$\Gamma\big(\mathcal{M}^{\text{weak}}, \mathcal{M}^{\text{strong}}\big) \quad \overset{H_0^{\text{strong}}}{\sim} \quad \chi^2_{(g-1)d_{freeinter}}, \tag{36}$$

where $d_{freeinter}$ is the number of free intercept parameters in the model structure.

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 (c.f. Kim, 2005).

While the LR test for MI is popular, scholars have taken issue with this approach (c.f. Drasgow et al., 2018). Their main argument is that with large sample sizes, rejection of the null is almost inevitable because the pre-specified model structure is likely not exactly the true model (e.g. Brannick, 1995; Kelloway, 1995). Of course, this is 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 detection methods discussed in the next section. The CFI of any model lies in the interval $[0, 1]$ and 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, the MI status of the items in a given model is generally unknown. 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 has several advantages over existing methods. For quick reference, Table 2 provides an overview of all detection methods that are discussed in this section.

Method	Basis	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	*original*	Obtains residuals from regressing each item on estimated latent variable. For each item, tests equal means of residuals across groups and tests for vanishing correlation between residuals and latent variable within each group. Items classified as non-invariant if means are different across groups or if there is non-vanishing correlation in at least one group.
R2	Residuals	*original*	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.*

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 sin-

gle 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. Second, as mentioned before, MI at the item level is primarily concerned with metric and scalar invariance. It 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 in a Bayesian framework. 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

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 at least in part be due to the fact that the literature is scattered across different fields of application. Moreover, the detection methods below have mostly been introduced in passing instead of being the focus of publications. As a result, they are also rarely formalized but more often just briefly described and considered as a step requiring the expertise of experienced researchers. To illustrate, an early way of detecting non-invariant items simply consisted in fitting a model as 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 and are introduced below.

All of the approaches try to identify violations of metric and scalar invariance simultaneously. As a result, they are not able to establish which type of MI is violated by a given item. Similar to global tests of MI, detection methods could also be designed to first detect violations of metric MI and only in case of absence of any violations move on to scalar MI at the item level. For the simulation study, this would however significantly increase the computational and cognitive demand, especially for the model-comparison based methods.[10] Because it is certainly more important to be made aware of a violation than to know

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

the exact nature of it, I limit the scope of the simulation study to the detection of the presence of any type of violation. Nonetheless, I test versions of the detection methods for violations of metric MI in the application section. The end of this section thus contains a brief description how the detection methods need to be altered to detect only violations of metric MI. I leave their rigorous introduction and the actual assessment of their performance in a simulation study to further research. Generally speaking, these versions could provide additional useful information for the purpose of model development. Notwithstanding, they are not the focus of this thesis and their introduction as well as their results in the application section should be viewed as an informal addition to the thesis.

A concern for all detection methods is the issue of multiple testing. During the development of my own method, it came naturally to include multiple testing corrections due to the high number of tests. Yet, the existing methods also conduct multiple tests that warrant a correction at least at the item-level. Notwithstanding, this issue isn't discussed in the original introductions of the detection methods and for the first two of the four existing methods, discussed below, a correction cannot even be applied. This is due to the fact that the non-invariance classification for these methods don't rely on classical statistical hypothesis testing on the basis of p-values. For the remaining two methods, however, I introduce both the original method and a Bonferroni-adjusted version below. Both are tested in the simulation study and will thus also provide an idea of the importance of multiple testing issues for the detection of non-invariant items.

4.1.1 Janssens (J)

First, an 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 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 all violations of MI come in the form of some groups deviating from others by having a loading equal to zero with respect to a given item, which is clearly a very restrictive assumption, resembling 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 have a loading that is close to the critical value of the sample distribution even if the differences across groups are virtually negligible (Cheung & Rensvold, 1999). Despite these theoretical concerns, I include the method in the simulation study because its computational demand is comparably low and it can serve as a baseline for performance of the detection methods.

Formally, for each item $i = 1, ..., p$ 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 $\wp_{i1}^l$ and $\wp_{i2}^l$ denote the corresponding p-values in group l, for example using a Wald test.[11] For a given significance level α, item i is said to violate scalar invariance if $\wp_{i1}^l < \alpha$ for at least one l, but at the same time $\wp_{i1}^{l'} \geq \alpha$ for at least one l'. The same goes for the loading parameters by considering the corresponding p-values. Evaluating violations of either type of MI, this approach yields as a set of identified items

$$S_{\text{J}} \coloneqq \left\{ i \middle| \wp_{i1}^l < \alpha \;\&\; \wp_{i1}^{l'} \geq \alpha \text{ or } \wp_{i2}^l < \alpha \;\&\; \wp_{i2}^{l'} \geq \alpha, \text{ for some } l, l' \right\}. \tag{37}$$

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 this requires the assumption that invariance holds for the parameters which remain constrained (Cheung & Rensvold, 1999).

To formalize, given structure $\mathcal{M}$, let $\mathcal{M}^{\text{strong}}$ denote the strongly constrained MGCFA model and $\mathcal{M}_i^{\text{strong}}$ a model which is identical except for the modification of lifting the equality constraints on the loadings and intercepts of item i. Since $\mathcal{M}^{\text{strong}}$ is nested in $\mathcal{M}_i^{\text{strong}}$, the MInd can be written as the LR statistic of these models

$$\Gamma\left(\mathcal{M}_i^{\text{strong}}, \mathcal{M}^{\text{strong}}\right) \overset{H_0}{\sim} \chi^2_{2(g-1)}. \tag{38}$$

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

$$S_{\text{MInd}} \coloneqq \left\{ i \;\middle|\; \Gamma\left(\mathcal{M}_i^{\text{strong}}, \mathcal{M}^{\text{strong}}\right) > t_\alpha \right\}, \tag{39}$$

which are all items for which the null hypothesis of no difference in goodness-of-fit is rejected as a result of lifting the corresponding constraints.

For a Bonferroni-adjusted version, where instead of the significance level α, the method uses a significance level of $\alpha^* = \alpha/p$, the set of non-invariant items $S_{\text{MInd}-\text{B}}$ identified by

[11] Which is what the `lavaan` package provides for each parameter.

this version can then be written as

$$S_{\text{MInd}-\text{B}} \coloneqq \left\{ i \;\middle|\; \Gamma\left(\mathcal{M}_i^{\text{strong}}, \mathcal{M}^{\text{strong}}\right) > t_{\alpha^*} \right\}. \tag{40}$$

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 by construction. 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 $\mathcal{M}^{\text{base}}$ as a fully unconstrained MGCFA model with model structure $\mathcal{M}$. This baseline model is then compared with several models for which the loading and intercept[12] of a single item $i = 1, ..., p$ is constrained to equality across groups while the remaining parameters remain free to vary across groups. This is repeated with changing reference items $j = 1, ..., p$, $j < i$. Taken together, these steps yield a procedure which involves one baseline model and $p(p-1)/2$ models that are all nested in the baseline model.

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

$$\Gamma_{ij} = \Gamma\left(\mathcal{M}^{\text{base}}, \mathcal{M}_{ij}^{\text{base}}\right) \overset{H_0}{\sim} \chi^2_{2(g-1)} \tag{41}$$

which can then be arranged in a strictly lower triangular matrix $\mathbf{\Gamma}$ of test statistics

$$\mathbf{\Gamma} \coloneqq \begin{bmatrix} & & \\ \Gamma_{21} & & \\ \vdots & \ddots & \\ \Gamma_{p1} & \dots & \Gamma_{p(p-1)} \end{bmatrix}. \tag{42}$$

According to Cheung and Rensvold's (1999) *triangle heuristic*, the set of non-invariant items can be identified as follows. First, for technical reasons, create a symmetric version of $\mathbf{\Gamma}$ by mirroring the lower triangle $(\mathbf{\Gamma} + \mathbf{\Gamma}^\mathsf{T})$. Second, simultaneously permute the rows and columns of this symmetric matrix with a suitable permutation matrix $\boldsymbol{P}$ to yield a matrix $\widetilde{\mathbf{\Gamma}} = \boldsymbol{P}(\mathbf{\Gamma} + \mathbf{\Gamma}^\mathsf{T})\boldsymbol{P}^\mathsf{T}$ which maximizes the number of consecutive rows counted from the

[12] Note 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.

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 the lower triangle of $\widetilde{\mathbf{\Gamma}}$. In other words, the permutation should rearrange the items such that significant statistics in the lower triangle appear in the lower rows of $\widetilde{\mathbf{\Gamma}}$. Cheung and Rensvold (1999) then consider those items which appear below the last row that contains no significant test statistics in the lower triangle to be non-invariant. More formally, let $\pi(i)$ be the position of item i in the permutation yielding $\widetilde{\mathbf{\Gamma}}$ and let $(p - m_\alpha)$ denote the number of invariant items as identified by the procedure, then the estimated set of non-invariant items S_{CR} of this procedure is given by

$$S_{\mathrm{CR}} \coloneqq \left\{ i \;\middle|\; \pi(i) \leq (p - m_\alpha) \right\}. \tag{43}$$

Note that $\widetilde{\mathbf{\Gamma}}$ 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 resulting sets are valid. They argue that the choice must be "made in light of substantive issues and underlying theory" (Cheung & Rensvold, 1999, p.12). 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. Further note that if all items are non-invariant, then the item corresponding to the first row in $\widetilde{\mathbf{\Gamma}}$ is still classified as invariant because the procedure only considers pairs of items where $i \neq j$.

In addition, I also implement a Bonferroni-corrected version of the CR approach, where instead of the significance level α, the method uses a significance level of $\alpha^* = \alpha/p$. As a result, the estimated set of non-invariant items $S_{\mathrm{CR-B}}$ of the Bonferroni-corrected version is given by

$$S_{\mathrm{CR-B}} \coloneqq \left\{ i \;\middle|\; \pi(i) \leq (p - m_{\alpha^*}) \right\}. \tag{44}$$

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 which is introduced in the appendix of this thesis. More specifically, the procedure works by first fitting a strongly constrained MGCFA model $\mathcal{M}^{\mathrm{strong}}$ of structure $\mathcal{M}$ as the baseline and determining its CFI. Additionally, for each $i = 1, ..., p$, a model $\mathcal{M}^{\mathrm{strong}}_{(-i)}$ is fitted where item i has been removed entirely from the 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 because of the equality constraints in

the baseline model. 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 MI. To summarize, the estimated set of non-invariant items S_{BV} of this procedure is given by

$$S_{\text{BV}} \coloneqq \left\{ i \;\middle|\; \text{CFI}\left(\mathcal{M}_{(-i)}^{\text{strong}}\right) \geq \text{CFI}\left(\mathcal{M}^{\text{strong}}\right) + 0.01 \right\}. \tag{45}$$

4.2 Novel Approach

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. More specifically, it is based on the residuals of these linear relationships. The use of residuals for diagnostic purposes in a CFA framework 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 model structure of the specified model.[13] Regardless, to the best of my knowledge, residual analysis has not been used for detecting non-invariant items. The original contribution of this thesis is therefore to devise such a method under partial MI by studying patterns in the residuals of the linear relationships of a given CFA model. I begin with an introduction of the fundamental logic of this method and prove that residuals can theoretically be used to identify non-invariance by laying out the residual behavior under MI or violations thereof. I then provide visual examples to further strengthen the intuition behind the idea. Finally, I introduce the actual implementation of the method and discuss how it can be further improved with a step-wise extension.

4.2.1 Residual Behavior

For the introduction, assume a single latent factor model with p items that load on the latent variable η. We further focus on a single item for which we want to determine the MI status. For the moment, we can thus omit the subscript i and denote the item Y. Regressing Y on η yields a regression intercept γ and slope β which allow us to write the residuals r in the regression as

$$r \coloneqq Y - (\gamma + \beta\eta). \tag{46}$$

Note that, by construction, the residuals of this regression (or any linear regression for that

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

matter) satisfy the following two properties:

$$\mathbb{E}[r] = 0 \;\&\tag{47a}$$

$$\mathrm{Cov}(r, \eta) = 0.\tag{47b}$$

They have expectation zero and are uncorrelated with the linear predictor of the regression. However, these two properties do not necessarily hold for groups within the data. To see this, the true DGP of the item must be allowed to vary across groups. In terms of MI, note that there are four different group-specific DGPs: A simultaneous violation of both metric and scalar MI, a violation of metric MI, a violation of scalar MI, and the case of MI. Let Y^l denote the group-specific item and η^l the group-specific latent variable for group $l = 1, ..., g$, where η^l is a random variable with expectation μ^l and variance ϕ^l. Explicitly writing out the four DGPs in this notation gives

- **Simultaneous violation of metric and scalar MI:**

$$Y^l = \tau^l + \lambda^l \eta^l + \varepsilon\tag{48}$$

  where $\tau^l \neq \tau^{l'}$ and $\lambda^k \neq \lambda^{k'}$ for some $l, l', k, k' \in \{1, ..., g\}$.
- **Violation of metric MI:**

$$Y^l = \tau + \lambda^l \eta^l + \varepsilon\tag{49}$$

  where $\lambda^l \neq \lambda^{l'}$ for some $l, l' \in \{1, ..., g\}$
- **Violation of scalar MI:**

$$Y^l = \tau^l + \lambda \eta^l + \varepsilon\tag{50}$$

  where $\tau^l \neq \tau^{l'}$ for some $l, l' \in \{1, ..., g\}$
- **Perfect MI:**

$$Y^l = \tau + \lambda \eta^l + \varepsilon\tag{51}$$

where, without loss of generality, residual invariance is assumed such that $\boldsymbol{\varepsilon}$ is not group-specific.

Given these group-specific DGPs, the residuals in group l can be written as

$$r^l := Y^l - \left(\gamma + \beta \eta^l\right)\tag{52}$$

where it is important to note that γ and β are still the intercept and slope of a linear regression of Y on η in the pooled data and not just in group l. In other words, the regression parameters are obtained while ignoring any (potential) grouping in the data.[14]

[14] To further clarify what that entails, it may be useful to think about how the pooled Y and η can be written in terms of their group-specific constituents. This can be done with the help of a multinomial random variable. For further details, refer to Section 8.3 in the appendix.

We have the following Lemma for the regression of these two variables.

Lemma 4.1. *Under perfect MI, regressing Y on η while ignoring the group structure yields intercept γ and slope β*

$$\gamma = \tau \tag{53a}$$

$$\beta = \lambda. \tag{53b}$$

Proof. The regression coefficient β from regressing Y on η is given by

$$\begin{aligned} \beta &= \frac{\text{Cov}\left(Y, \eta\right)}{\text{Var}\left(\eta\right)} \overset{(51)}{=} \frac{\text{Cov}\left(\tau + \lambda\eta + \varepsilon, \eta\right)}{\text{Var}\left(\eta\right)} \\ &= \frac{\text{Cov}\left(\lambda\eta, \eta\right)}{\text{Var}\left(\eta\right)} \\ &= \lambda \end{aligned} \tag{54}$$

and the intercept γ by

$$\begin{aligned} \gamma &= \mathbb{E}\left[Y\right] - \beta\mathbb{E}\left[\eta\right] \overset{(51)}{=} \mathbb{E}\left[\tau + \lambda\eta + \varepsilon\right] - \beta\mathbb{E}\left[\eta\right] \\ &= \tau + \lambda\mathbb{E}\left[\eta\right] - \beta\mathbb{E}\left[\eta\right] \overset{(54)}{=} \tau + \lambda\mathbb{E}\left[\eta\right] - \lambda\mathbb{E}\left[\eta\right] \\ &= \tau. \end{aligned} \tag{55}$$

□

The fundamental idea of the new method is that the result of Lemma 4.1 cannot apply under any violation of metric or scalar MI for all group-specific CFA loadings and intercepts. The intuition can be reduced very simply to this: We conduct only a single regression of Y on η, yielding a single intercept and slope. At the same time, the nature of a violation of metric or scalar MI is that at least two groups have a different intercept or loading. As a result, under an MI violation, there are at least three parameters relevant for the true DGP at the pooled level. For example, suppose a two-group setting with groups A and B and a pure scalar MI violation. The DGP thus contains two group-specific intercept parameters, τ^A and τ^B, and the loading λ which is shared by both groups. Even if in the regression, $\beta = \lambda$ (which is not true in general), it is still impossible for $\gamma = \tau^A = \tau^B$ because $\tau^A \neq \tau^B$. In other words, for some or all groups, regression intercept and slope will not be equal to their group-specific intercept and loading in the true DGP. The exact difference between the group-specific parameters in the DGP and the regression parameters can be computed and I derive them in the appendix for further illustration. However, they are not necessary for the remainder of this section. Instead, it is important to realize that this discrepancy results in systematic patterns with regard to the expectation of the residuals and their covariance with the latent variable for some or all of the groups. To formalize this, the behavior of the residuals under different violations of MI can be condensed in the following theorem.

Theorem 4.2. *Metric and scalar MI are satisfied if and only if*

$$\mathbb{E}\left[r^l\right] = 0, \textit{ and} \tag{56a}$$

$$\mathrm{Cov}\left(r^l, \eta^l\right) = 0 \tag{56b}$$

for all $l = 1, ..., g$.

Proof. We begin with the case of a simultaneous violation of metric and scalar MI.

$$\begin{aligned} \mathrm{Cov}\left(r^l, \eta^l\right) &\overset{(48)}{=} \mathrm{Cov}\left(\tau^l + \lambda^l\eta^l + \varepsilon - \gamma - \beta\eta^l, \eta^l\right) \\ &= \mathrm{Cov}\left(\lambda^l\eta^l - \beta\eta^l, \eta^l\right) \\ &= \left(\lambda^l - \beta\right)\mathrm{Var}\left(\eta^l\right). \end{aligned} \tag{57}$$

Because $\mathrm{Var}\left(\eta^l\right) > 0$,

$$\mathrm{Cov}\left(r^l, \eta^l\right) = 0 \iff \lambda^l = \beta. \tag{58}$$

For this to hold in all groups, we require that $\beta = \lambda^1 = \lambda^2 = ... = \lambda^g$ which contradicts the assumed metric violation.

The same argument can be made in the presence of a sole violation of metric MI. Formally,

$$\begin{aligned} \mathrm{Cov}\left(r^l, \eta^l\right) &\overset{(49)}{=} \mathrm{Cov}\left(\tau + \lambda^l\eta^l + \varepsilon - \gamma - \beta\eta^l, \eta^l\right) \\ &= \left(\lambda^l - \beta\right)\mathrm{Var}\left(\eta^l\right). \end{aligned} \tag{59}$$

So the same logic that β cannot be equal to λ^l for all l applies.

Next, consider the case of a violation of scalar MI. We have

$$\begin{aligned} \mathbb{E}\left[r^l\right] &\overset{(50)}{=} \mathbb{E}\left[\tau^l + \lambda\eta^l + \varepsilon - \gamma - \beta\eta^l\right] \\ &= \tau^l - \gamma + (\lambda - \beta)\mu^l. \end{aligned} \tag{60}$$

If $\mu^l = 0$ for all l, we can argue analogously as for the loadings above that

$$\mathbb{E}\left[r^l\right] = \tau^l - \gamma = 0 \iff \tau^l = \gamma \tag{61}$$

which contradicts the violation of scalar MI.

If $\mu^l \neq 0$ for some or all l, $\mathbb{E}\left[r^l\right] = 0$ if and only if

$$\tau^l - \gamma = (\beta - \lambda)\mu^l. \tag{62}$$

In other words, the difference between the regression intercept and the group-specific intercept in the DGP may be cancelled out and the expected residual may be equal to zero.

Suppose there was some constellation of all group-specific intercepts and all latent variable expectations that ensured that $\mathbb{E}\left[r^l\right] = 0$ for all l then this implies that

$$(\beta - \lambda) \neq 0 \iff \beta \neq \lambda. \tag{63}$$

In turn, this yields the following result:

$$\begin{aligned} \text{Cov}\left(r^l, \eta^l\right) &\overset{(50)}{=} \text{Cov}\left(\tau^l + \lambda\eta^l + \varepsilon - \gamma - \beta\eta^l, \eta^l\right) \\ &= \text{Cov}\left((\lambda - \beta)\eta^l, \eta^l\right) \\ &= (\lambda - \beta)\text{Var}\left(\eta^l\right) \neq 0 \end{aligned} \tag{64}$$

because $\text{Var}\left(\eta^l\right) > 0$.

In summary, the steps above conclude half of the iff statement in the theorem and prove that any violation of MI results in the group-specific residuals having non-zero expectation or non-vanishing correlation with the latent variable. To conclude the proof, we further need to show that under perfect MI, the two components hold. We have that

$$\begin{aligned} \mathbb{E}\left[r^l\right] &\overset{(51)}{=} \mathbb{E}\left[\tau + \lambda\eta^l + \varepsilon - \gamma - \beta\eta^l\right] \\ &= \tau + \lambda\mu^l - \gamma - \beta\mu^l \end{aligned} \tag{65}$$

In accordance with Lemma 4.1, we can replace $\gamma = \tau$ and $\beta = \lambda$ and as a result it is easy to see that

$$\mathbb{E}\left[r^l\right] = \tau - \tau + \lambda\eta^l - \lambda\eta^l = 0. \tag{66}$$

Similarly, we have

$$\begin{aligned} \text{Cov}\left(r^l, \eta^l\right) &\overset{(51)}{=} \text{Cov}\left(\tau + \lambda\eta^l + \varepsilon - \gamma - \beta\eta^l, \eta^l\right) \\ &= \text{Cov}\left(\varepsilon, \eta^l\right) \\ &= 0, \end{aligned} \tag{67}$$

which completes the second half of the iff statement and thus concludes the proof. □

4.2.2 Illustrating Residual Behavior

To further illustrate the implications of Theorem 4.2, Figure 3 visualizes an example of the patterns in the residuals of a single item for two groups across the four DGPs. In all panels, the groups differ in their true latent mean with the red group scoring higher than the black group. 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 as thick solid lines in the figure. Panel A corresponds to perfect MI where the two groups originate from identical DGPs. The fitted lines show that both groups have a mean close to zero and that there is no correlation with the latent variable in either group. In panel B, scalar invariance is violated and the red group's intercept is shifted up by 0.9 units compared to the black group. 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 compared to the black group. Finally, panel D combines both violations using the same values. In both panels, the correlation with the latent variable in the black group is very pronounced. Generally, across the panels, it is easy to see that the two criteria of zero mean and vanishing correlation in the groups only hold in the MI setting shown in panel A. For any violation of MI, panels B-D show that the residuals exhibit deviations from at least one of these criteria in at least one of the two groups.

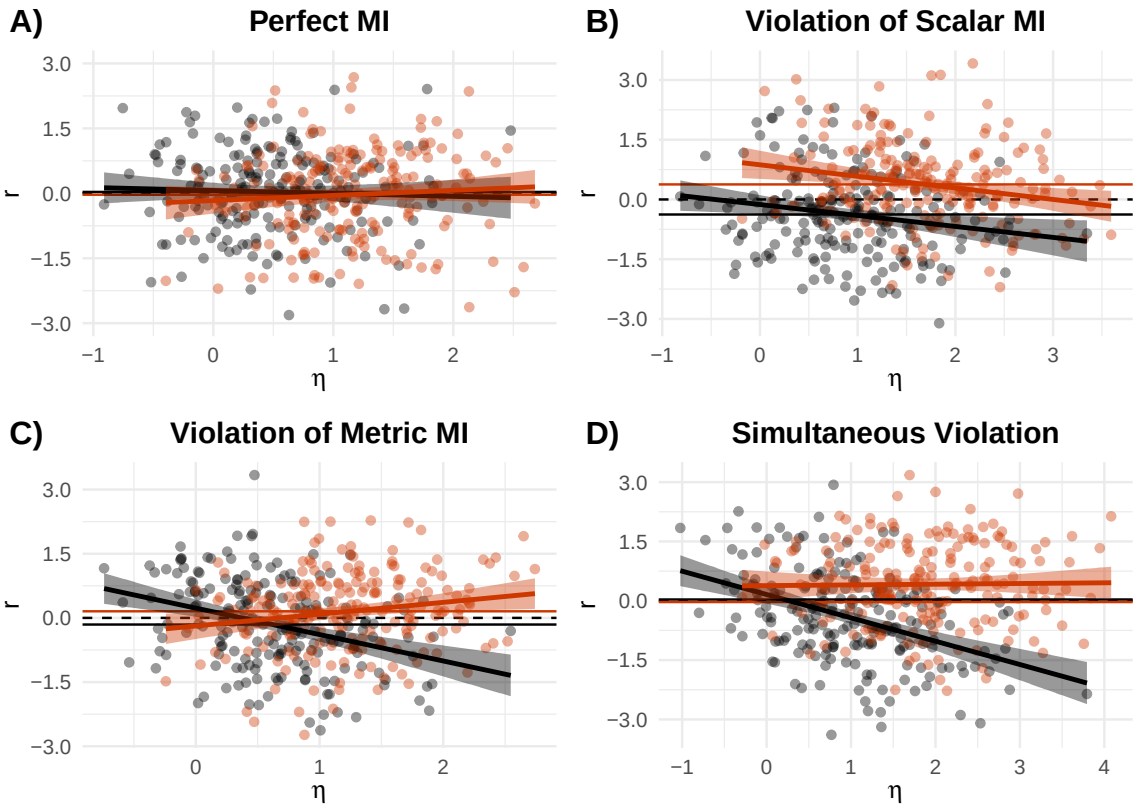

Figure 3: *Illustration of the residuals from four DGPs across two groups (black and red). Thick solid lines are regression lines with 95-% confidence bands. Thin horizontal lines are group means.*

Furthermore, the example shows that it's possible to visualize the MI status of the items in CFA models, which may be helpful for applied researchers. The tools 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.

4.2.3 Implementation as a Detection Method (R1)

To devise a detection method from these ideas, the two components of across-group comparisons of the residual means and correlations between the residuals and the latent variable within each group must be formalized for hypothesis testing. 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.

Before going into details, it is important to note the obvious obstacle for such a method. The key problem is the very raison d'être of CFA itself: the latent variable η is unknown. Therefore, we cannot really regress Y on η. The obvious - albeit naive - solution is to rely on estimates $\hat{\eta}$ instead which has the conspicuous shortcoming that these estimates are problematic under non-invariance. Unfortunately, there is no other way around this issue in a world in which latent variables exist. It is the price to pay for the simplicity of the novel approach. The implication of this issue is that we at least require a setting of partial MI such that the estimates at least resemble the "true" latent variables. I return to this issue and the amelioration of its consequences with the implementation of a step-wise version of the detection method below.

In the following, we turn to sample versions and estimates, not just of the latent variable, but also of the regression parameters, the residuals, etc. Moreover, we again consider all $i = 1, ..., p$ items for which we want to test for MI. We can thus write the residuals in group l as

$$\hat{r}_i^l = Y_i^l - \hat{\gamma}_i - \hat{\beta}_i \hat{\eta}^l \tag{68}$$

where $\hat{\gamma}_i$ and $\hat{\beta}_i$ are the estimated parameters of the linear regression of Y_i on $\hat{\eta}$ with the pooled data.

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

$$\nu_i^l := \mathbb{E}\left[\hat{r}_i^l\right] \tag{69}$$

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

$$H_0 : \nu_i^1 = ... = \nu_i^g, \tag{70}$$

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

$$\frac{\frac{1}{(g-1)}\sum_{l=1}^{g} n^l \left(\hat{\nu}_i^l - \hat{\nu}_i\right)^2}{\frac{1}{(N-g)}\sum_{l=1}^{g}\sum_{j=1}^{n^l}\left(\hat{r}_{ij}^l - \hat{\nu}_i^l\right)^2} \overset{H_0}{\sim} F_{(g-1),(N-g)}, \tag{71}$$

where $\hat{\nu}_i^l$ and $\hat{\nu}_i$ are the sample means of the residuals in group l and in the full sample of size N, respectively. Note that the latter sample mean is zero by construction. However, it is kept in equation (71) to emphasize that the statistic is the standard ANOVA F-test. For a more detailed account, we refer to an introductory ANOVA book, e.g. Oehlert (2000).

With regard to the second component of vanishing correlation between the residuals and the latent variable, the linear relationship can be tested via the corresponding coefficients in separate linear regressions of $\hat{r}_i^l$ on $\hat{\eta}^l$ for each item and each group.[15] It is important to stress that this is a secondary regression after already having obtained the residuals from a first regression. In total, the method entails a total of pg secondary regressions. Let κ_i^l and ω_i^l denote the regression intercepts and slopes of the secondary regressions. The null hypothesis of no correlation can then be written as

$$H_0 : \omega_i^l = 0. \tag{72}$$

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

$$\frac{\hat{\omega}_i^l}{\text{se}\left(\hat{\omega}_i^l\right)} \overset{H_0}{\sim} t_{n^l-2} \tag{73}$$

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

$$\hat{\text{se}}\left(\hat{\omega}_i^l\right) = \sqrt{\frac{\sum_{j=1}^{n^l}\left(\hat{r}_{ij}^l - \hat{\kappa}_i^l - \hat{\omega}_i^l\hat{\eta}_j^l\right)^2}{(n^l-2)\sum_{j=1}^{n^l}\left(\hat{\eta}_j^l - \frac{1}{n^l}\sum_{j=1}^{n^l}\hat{\eta}_j^l\right)^2}}. \tag{74}$$

Again, details can be found in any standard introductory statistics textbook introducing linear regression, e.g. Fahrmeir et al. (2013).

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

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

that

$$\widetilde{\wp}_i \coloneqq (g+1)\min(\wp_{i,1}, ..., \wp_{i,g+1}), \tag{75}$$

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:

$$S_{\mathrm{R1}} \coloneqq \left\{ i \;\middle|\; (p - \pi(i) + 1)\widetilde{\wp}_i < \alpha \right\}, \tag{76}$$

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

4.2.4 Step-wise Version (R2)

As noted previously, 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 existing methods 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 of structure $\mathcal{M}$, 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, 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 to 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 of structure $\mathcal{M}$ can best be described by the following algorithm:

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

(c) Apply R1 to $\mathcal{M}_{t+1}$ and update $\widetilde{\wp}$

(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.[16] Every step within the CFA framework, i.e. model fitting, testing, etc., is executed using the `lavaan` package (Rosseel, 2012; v0.6-9).

4.4 Detecting Items Violating Metric MI

As discussed above, the detection methods do not distinguish between the type of MI, i.e. metric or scalar, that is violated by any given item. Nonetheless, it is relatively straightforward to implement versions of the detection methods, that only consider violations of metric MI, i.e. the weak requirement for MI. In general, the expectation for these methods is that they classify a subset of the items that are detected by the standard versions.

In the following, I briefly describe for each method how it needs to be altered in order to focus on the invariance of loadings while disregarding the intercepts. All other steps of the methods remain unchanged compared to their original version. For the R approach, I go into greater detail by proving that the residuals enable the drawing of conclusions about the type of violation. Note that I exclude the J approach from the application section because its theoretical weaknesses were corroborated by the results of the simulation study.

4.4.1 Modification Indices (MInd)

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 constraints for each item. The metric version works identically, but only constrains the loadings for the baseline model and modification indices refer to the lifting of single loading constraints. Therefore, the degrees of freedom of the LR test also need to be adjusted accordingly. As a result, the metric version completely disregards the intercepts and lets them vary freely across groups in the baseline model as well as the comparison models.

[16] 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).

4.4.2 Cheung & Rensvold (CR)

Similarly, for the metric version of the CR approach, in the construction of $\mathcal{M}_{ij}^{\text{base}}$, the equality constraints are only imposed on the loading parameter of item i instead of both the loading and intercept. As with the metric MInd method, the degrees of freedom in the LR test are also adjusted accordingly. Recall that Cheung and Rensvold (1999) are agnostic about the intercepts. Consequently, this may even be their originally intended method.

4.4.3 Byrne & Van de Vijver (BV)

For the BV method, a metric version can be obtained by changing the baseline model from a strongly constrained to a weakly constrained MGCFA model $\mathcal{M}^{\text{weak}}$. Thus, the intercepts are free to vary across groups in the baseline model. The metric BV method still removes items entirely, but the removal of an item will not result in an increase in CFI due to violations of invariance of the intercepts because they are free to vary in the baseline model.

4.4.4 Rieger (R1 & R2)

For the R approaches, the necessary changes aren't quite as obvious. It may be tempting to simply choose the correlation component as relating to loadings, but it is clear from panel B of Figure 3 that non-zero correlation can result under scalar invariance if the true latent means vary across groups. Instead, the following Theorem can provide further insight. To omit some indices for the sake of clarity, we again consider a single item Y as in Theorem 4.2.

Theorem 4.3. *Metric MI is satisfied if and only if*

$$\omega^l = \omega^k \tag{77}$$

for all pairs $l, k \in \{1, ..., g\}$.

Proof. First, recall that ω^l, the regression slope in the secondary regressions of item Y's residuals on the latent variable in group l, is given by

$$\omega^l = \frac{\text{Cov}\left(r^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)}. \tag{78}$$

Further note that in the four relevant DGPs, metric MI is satisfied in both the case of a pure scalar MI violation and in the case of perfect MI, shown in equations (50) and (51), respectively. We thus need to show that the statement of the theorem holds in these cases. For the case of perfect MI, the statement is true by Theorem 4.2 which shows that the

covariance in the numerator of equation (78) is equal to zero and thus also

$$\omega^l = 0 \; \forall \; l. \tag{79}$$

For the scalar MI violation, we have

$$\begin{aligned} \frac{\text{Cov}\left(r^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} &\overset{(50)}{=} \frac{\text{Cov}\left(\tau^l + \lambda\eta^l + \varepsilon - \alpha - \beta\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= (\lambda - \beta)\frac{\text{Cov}\left(\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= (\lambda - \beta). \end{aligned} \tag{80}$$

which no longer depends on l.

Next, we need to show that the statement doesn't hold if metric MI is violated, i.e. for the remaining two cases in equations (48) and (49). For the former, we have

$$\begin{aligned} \frac{\text{Cov}\left(r^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} &\overset{(48)}{=} \frac{\text{Cov}\left(\tau^l + \lambda^l\eta^l + \varepsilon - \alpha - \beta\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= (\lambda^l - \beta)\frac{\text{Cov}\left(\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= \lambda^l - \beta \end{aligned} \tag{81}$$

and for the latter, we have the same result by

$$\begin{aligned} \frac{\text{Cov}\left(r^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} &\overset{(49)}{=} \frac{\text{Cov}\left(\tau + \lambda^l\eta^l + \varepsilon - \alpha - \beta\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= (\lambda^l - \beta)\frac{\text{Cov}\left(\eta^l, \eta^l\right)}{\text{Var}\left(\eta^l\right)} \\ &= \lambda^l - \beta. \end{aligned} \tag{82}$$

In a similar argument to the one made in Theorem 4.2,

$$\lambda^l - \beta = \lambda^k - \beta \; \forall \; l, k \iff \lambda^1 = ... = \lambda^g, \tag{83}$$

which contradicts the presence of a violation of metric invariance. □

In other words, Theorem 4.3 shows that in the secondary regressions of $\hat{r}_i^l$ on η^l, groups have parallel regression lines if and only if metric MI is satisfied. An example of this can also be seen in panel B of Figure 3.

Thus, for a metric version of the R approach, what is of relevance is that the regression coefficients differ across groups. This can be tested by adding an interaction term to the regression of the residuals on the latent variable estimates. More specifically, for the met-

ric version, the residuals of all groups are regressed on the latent variable estimates, the group membership and the interaction of those two covariates. If the interaction term explains a significant share of the variance in the residuals, this must be due to differences in the group-specific regression coefficients of the latent variable. Formally, this test is implemented as an F-test of the group membership variable. Since the step-wise R2 method internally builds on R1, a metric version can be obtained by simply using the metric R1 version for each step.

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 DGP 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 DGP 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 MI. 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). Including relatively more non-invariant items leads to numerous convergence issues when trying to fit even the base models. The DGP thus maintains a relatively strong partial MI setting, where more than half of the items are indeed measurement invariant. In total, the unique combinations of parameter values yield 896 different simulation settings.

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

$$\mu^l \sim \mathcal{N}\left(0, 0.3^2\right). \tag{84}$$

The g true standard deviations $\sqrt{\phi}$ are the absolute value of normal distribution draws with mean one and a standard deviation of 0.1, i.e.

$$\sqrt{\phi^l} \sim \mathcal{N}^+\left(1, 0.1^2\right). \tag{85}$$

These parameter samples are then used to sample positions on the latent variable $\boldsymbol{\eta}^l$ for each observation $j = 1, ..., n$ within each group. Note that n is the number of observations in each group such that there is a total of $N = gn$ observations with all groups having equal size. The observation index j is nested in the group index l, denoted $l(j)$. The latent

variables are then sampled as

$$\eta^{l(j)} \sim \mathcal{N}\left(\mu^l, \phi^l\right). \tag{86}$$

Turning to the items, intercepts $\boldsymbol{\tau}$ and loadings $\boldsymbol{\lambda}$ for each of the $i = 1, ..., p$ items Y_i are sampled. The loadings come from a normal distribution with mean 0 and a standard deviation of 0.5, i.e.

$$\tau_i \sim \mathcal{N}\left(0, 0.5^2\right) \tag{87}$$

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

$$\lambda_i \sim \text{Unif}(0.65, 0.85). \tag{88}$$

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

$$Y_i^{l(j)} \sim \mathcal{N}\left(\tau_i + \lambda_i \eta^{l(j)}, \left(1 - \lambda_i^2\right)^2\right) \tag{89}$$

Note that if it stopped here, the DGP would reflect perfect MI. To create a setting of partial MI, a total of hg groups and m items are 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_τ and δ_λ, respectively, where $\delta_\tau, \delta_\lambda \in \{0, 0.2\}$. As a result, there are four different settings with regard to the type of (violation of) MI:

1. metric and scalar non-invariance ($\delta_\tau = \delta_\lambda = 0.2$)
2. purely scalar non-invariance ($\delta_\tau = 0.2$ & $\delta_\lambda = 0$)
3. purely metric non-invariance ($\delta_\tau = 0$ & $\delta_\lambda = 0.2$)
4. MI ($\delta_\tau = \delta_\lambda = 0$).

Additionally, the sign of each bias is randomly sampled for each group and item. Let l' and i' denote a group and an item that were sampled to be affected by non-invariance. Then

$$Y_{i'}^{l'(j)} \sim \mathcal{N}\left(\tau_{i'} \pm \delta_\tau + (\lambda_{i'} \pm \delta_\lambda)\eta^{l'(j)}, \left(1 - (\lambda_{i'} \pm \delta_\lambda)^2\right)^2\right). \tag{90}$$

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 (c.f. Pokropek et al., 2019). 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 items	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 items	k	$\{1, 2, 3\}$	
Magnitude of bias on intercepts	δ_τ	$\{0, 0.2\}$	sign of bias randomly and independently sampled for each group and item
Magnitude of bias on loadings	δ_λ	$\{0, 0.2\}$	———— 〃 ————

Table 3: *Simulation parameters.*

5.2 Simulation Setup

For each of the 896 unique simulation parameter value combinations, I simulate 100 datasets according to the procedure introduced above, resulting in $89{,}600$ total iterations. In each iteration, all detection methods are employed to detect non-invariant items. This is done with the standard versions of the detection methods that do not attempt to distinguish between the type of MI that is violated, but simply classify the items which violate either type. 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. Analysis of their performance is conducted at the item-level. In total, there are $436{,}800$ items across all 896 parameter combinations. Of those, $109{,}200$ are non-invariant.

Given the nature of the output of each method and that the truly non-invariant items are known, confusion matrices and derived metrics for evaluating the performance of each method under the different simulation parameter specifications can be generated. 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.*

The main performance metrics for the simulation study are the *sensitivity* (true positive rate) and *Specificity* (true negative rate) of the detection methods. Sensitivity is defined as the share of correctly identified non-invariant items (TP) among the truly non-invariant items (TP + FN):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{91}$$

Specificity is defined as the share of correctly identified invariant items (TN) among all truly invariant items (TN + FP):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{92}$$

A good detection method should have both high sensitivity and specificity. A detection method which has high sensitivity, but low specificity is useless because it is unclear whether its positive classifications are valid or not. In the extreme, it would be easy to maximize sensitivity with a method that indiscriminately returns all positive predictions. On the other hand, a method with high specificity and low sensitivity may miss a lot of the relevant non-invariant items. However, as long as one is aware of the low sensitivity, the method may still be useful in providing some guidance as to where to start with model improvement. In that case, it is crucial that global tests of MI are conducted along the way.

5.3 Results

The remainder of this section includes the key findings from the simulation study. Before going into detail with regard 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 or in visual form in Figure 4 which highlights the trade-off between sensitivity and specificity. They include the aggregate sensitivity and specificity under a simultaneous violation of both metric and scalar MI ($\delta_\tau = \delta_\lambda = 0.2$), violation of scalar MI ($\delta_\tau = 0.2$ & $\delta_\lambda = 0$), violation of metric MI ($\delta_\tau = 0$ & $\delta_\lambda = 0.2$), as well as in the case of perfect MI ($\delta_\tau = \delta_\lambda = 0$). The subsequent detailed analysis then focuses on the extremes, i.e. simultaneous metric and scalar non-invariance as well as perfect MI.

Starting with the case of simultaneous metric and scalar non-invariance, all methods but the J approach perform well in terms of sensitivity. All other methods detected at least 70% of the items that were truly non-invariant with most methods ranging between 80% and 100%. The highest sensitivity is achieved by the MInd approach, with less than 3% of non-invariant items that were not detected. Yet, this high sensitivity comes at the cost of the lowest specificity, even though the Bonferroni correction results in a considerable mitigation of this weakness. For the CR approach, the Bonferroni correction is less effective and both versions perform almost identically across all different types of MI violations. Either way, the MInd, CR, and R1 approaches classify too many items as non-invariant,

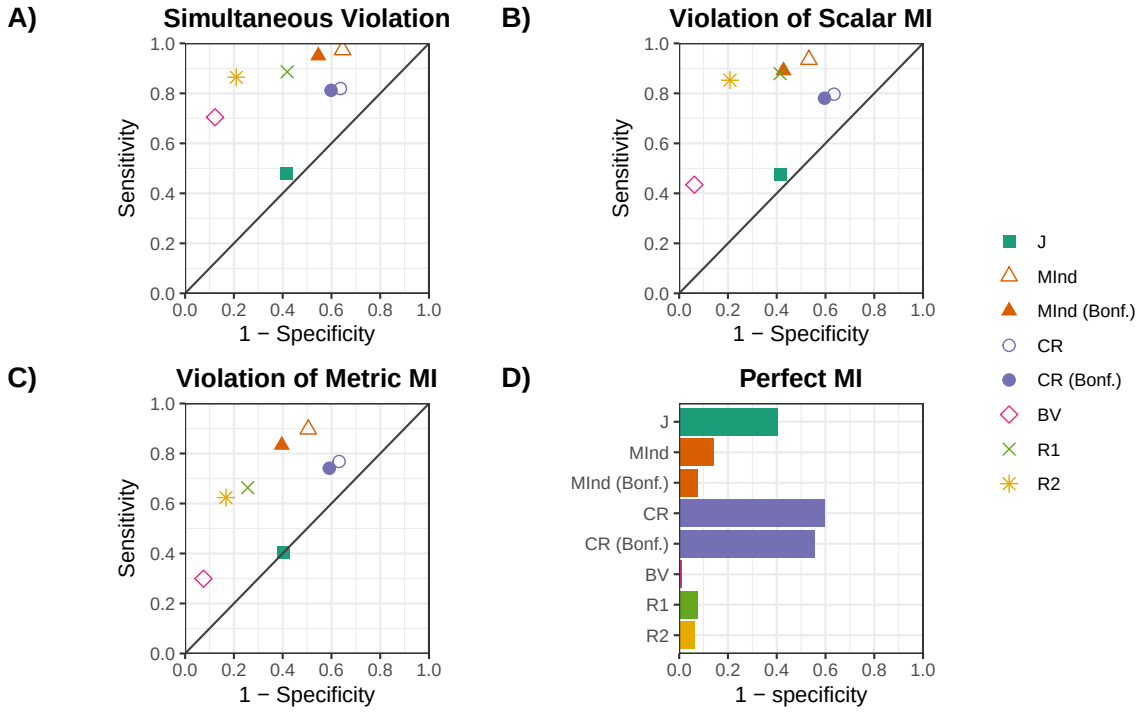

Figure 4: *Aggregate performance of the detection methods across all simulation settings.*

		$\delta_\tau > 0$				$\delta_\tau = 0$		
		$\delta_\lambda > 0$		$\delta_\lambda = 0$		$\delta_\lambda > 0$		$\delta_\lambda = 0$
		simultaneous		scalar		metric		perfect MI
	Method	**Sens**	**Spec**	**Sens**	**Spec**	**Sens**	**Spec**	**Spec**
■	J	0.478	0.584	0.476	0.585	0.402	0.596	0.595
△	MInd	0.973	0.354	0.935	0.468	0.897	0.495	0.857
▲	MInd (Bonf.)	0.951	0.454	0.891	0.572	0.834	0.604	0.923
○	CR	0.819	0.363	0.797	0.365	0.768	0.369	0.401
●	CR (Bonf.)	0.812	0.401	0.780	0.404	0.741	0.409	0.446
◇	BV	0.704	0.877	0.435	0.938	0.299	0.925	0.990
×	R1	0.886	0.582	0.879	0.586	0.663	0.743	0.925
✳	R2	0.864	0.790	0.853	0.792	0.624	0.833	0.937

Table 5: *Aggregate performance of the detection methods across all simulation settings. Each performance metric is based on the classification of* 109,200 *items. In the setting of a simultaneous metric and scalar violation, a total of* 18 *items were not classified by the J and BV approach due to computational errors, e.g. convergence issues. Further 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.*

resulting in a high number of false positives. Only the BV and R2 approaches achieve satisfactory specificity of around 88% and 79%, respectively. In terms of the trade-off with sensitivity, the R2 approach performs best with a fairly high sensitivity and the second best specificity. These trade-offs are particularly obvious when referring to Panel A in Figure 4. Further comparing the R2 with the R1 approach, it appears that the step-wise approach yields a substantial increase in specificity at the cost of a negligible decrease in sensitivity.

This holds true for all types of MI violations in the simulation study.

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 about 27 percentage points compared to the simultaneous non-invariance case.

For the metric non-invariance setting, the results seem to suggest that this is the most difficult violation to detect. In general, all methods perform even worse than in the pure scalar non-invariance case. Particularly the sensitivity of the BV and both R approaches declines significantly compared to the simultaneous violation of MI. Despite the decrease in sensitivity of about 24 percentage points, the R2 approach remains the best performing method under the trade-off between sensitivity and specificity. It still correctly classified 62% of the truly non-invariant items and 83% of truly invariant items. Nonetheless, its advantage over the Bonferroni-adjusted MInd approach is only minor in this setting.

In the perfect MI case, 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 by construction because there are no 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 MI violations, superfluous detection is very rare. Consequently, there really is no reason not to use these detection methods. In the best case scenario, i.e. when they aren't needed, they come at virtually no cost. On the other hand, this reasoning suggests that the use of the J and CR approaches is not advisable.

Before moving on to the more detailed analysis, two points are noteworthy. First, the J approach is the only method which consistently performs very 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 and show that it cannot be considered a useful method for detecting non-invariant items under any setting. For completeness' sake, I keep the J approach in the figures below, but mostly disregard it in the interpretation of the results. Second, the aggregate results above show that the Bonferroni corrections for the MInd and CR approaches are generally useful. In the trade-off between sensitivity and specificity, they generally improve the methods' specificity at a cost of a fraction of reduced sensitivity. I thus only consider the Bonferroni-corrected versions of these methods for the remainder of this analysis.

5.3.1 Sensitivity

Figure 5 sheds some more light on how sensitivity is affected by the number of observations per group n, the number of items p, and the number of non-invariant items m.

Recall that in the following, all comparisons and figures refer to the case where both the intercepts and loadings are simultaneously non-invariant. All other simulation parameters were ignored, so the overall levels of sensitivity reflect an average of their parameter values. Further note that the confidence intervals in this and the remaining figures are vanishingly small due to the high number of classified items.

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 non-invariant items increases. The BV approach seems to be affected most by increases in m and for $m > 1$ also no longer benefits 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 is sufficiently high: If n is at least 500, all of these methods are able to detect nearly 100% of truly non-invariant items.

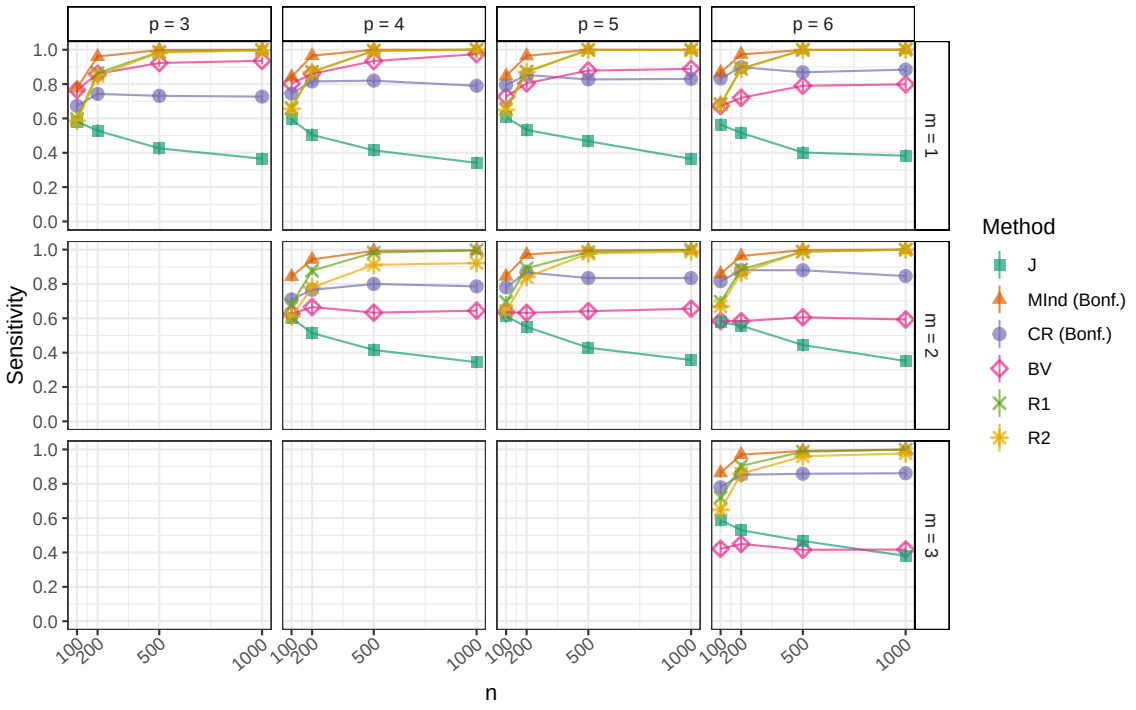

Figure 5: *Sensitivity of different detection methods as a function of n, p, and m under simultaneous metric and scalar non-invariance. Vertical lines represent 95% Clopper-Pearson confidence intervals.*

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 6 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 $h = 0.5$.[17] 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 30 percentage points when two instead of just one of four groups are affected by non-invariance. These findings as well as the ones above with regard to the BV method show that its overall mediocre sensitivity can be explained quite well by a few simulation settings that affect it more than others.

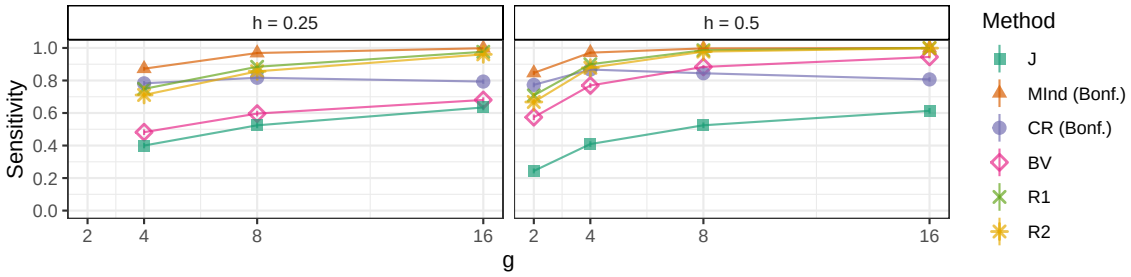

Figure 6: *Sensitivity of different detection methods as a function of g and h under simultaneous metric and scalar non-invariance. Vertical lines represent 95% Clopper-Pearson confidence intervals.*

5.3.2 Specificity

Figures 7 and 8 were created analogously to Figures 5 and 6, 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 7 shows that the vast majority of methods tends to detect fewer true negatives as n increases. Put differently, the number of false rejections of the null hypothesis of MI increases with n. Additionally, they tend to reach higher specificity with more items in the model and fewer items affected by non-invariance. However, not all methods are affected in the same way by this. As long as there are more than three items, the best results are by far achieved with the BV method which is also robust to varying group 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 under-identified model which 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: The R1 approach for small group sizes and the R2 approach across the board. The R1 approach is very much affected by increases in n while the specificity of the R2 approach only decreases slightly as a result of increasing group size. For the MInd and R1 approaches, the results indicate very poor performance for larger group sizes and show that they are particularly susceptible to changes in sample size: For the largest sample size of 1000, specificity of the MInd method almost drops to 0, meaning that it effectively classifies every item as being

[17] When comparing the two panels, note that the left panel doesn't include any cases for $g = 2$ because $g = 2$ times $h = 0.25$ isn't a natural number.

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 than the MInd and R1 approach.

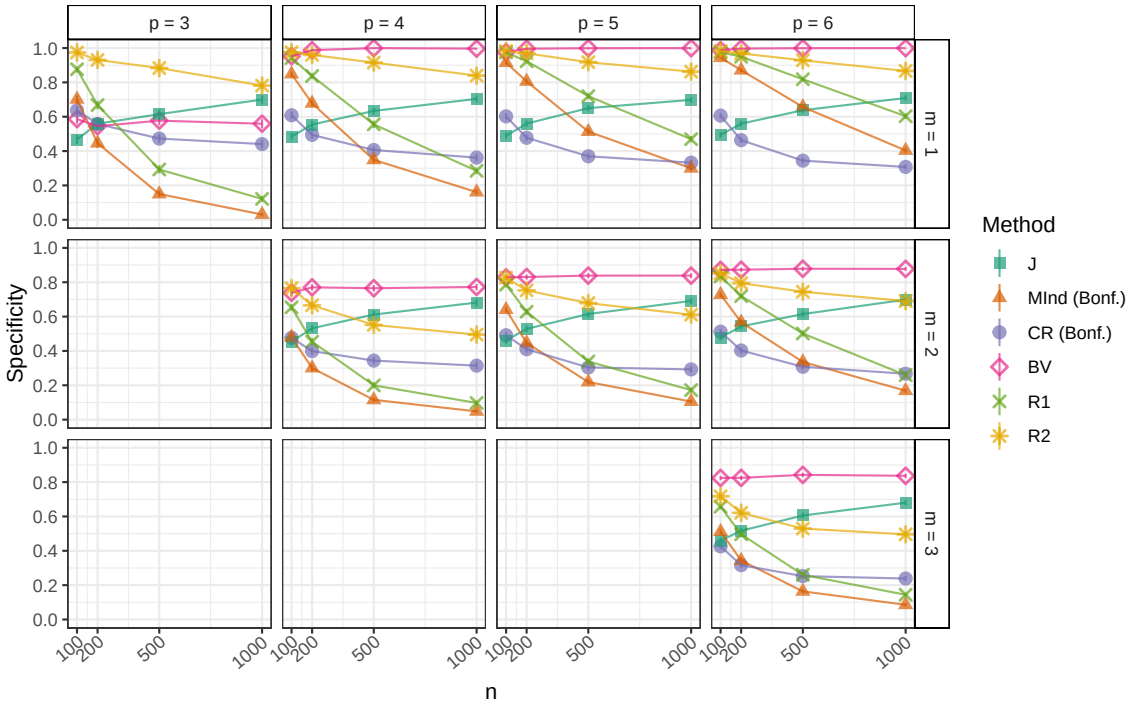

Figure 7: *Specificity of different detection methods as a function of n, p, and m under simultaneous metric and scalar non-invariance. Vertical lines represent 95% Clopper-Pearson confidence intervals.*

The same applies to increases in the number of groups, as Figure 8 shows: As g increases, specificity tends to decline, albeit slowly. Only the BV method resists this effect and exhibits an almost horizontal line, indicating that it's specificity remains relatively constant for the different settings of g. Otherwise, the remaining methods are similarly affected and the general ranking of methods still applies. 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 8.[18] 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 and J approaches achieve (slightly) higher specificity in the balanced setting while the effect for the R2 approach depends on g.

[18]However, when comparing the two panels, again note that it doesn't include any cases for $g = 2$.

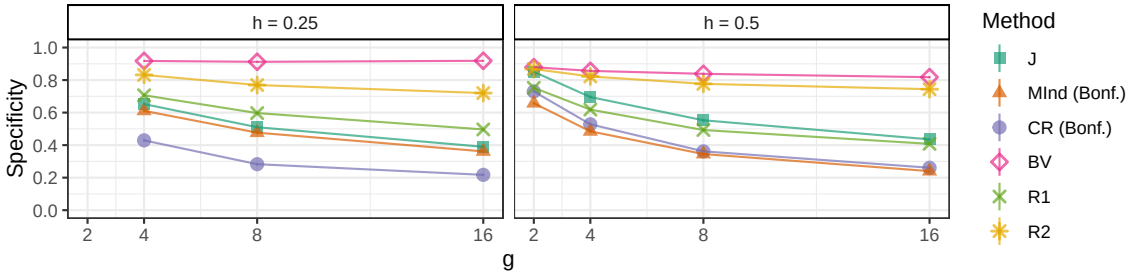

Figure 8: *Specificity of different detection methods as a function of g and h under simultaneous metric and scalar non-invariance. Vertical lines represent 95% Clopper-Pearson confidence intervals.*

5.3.3 Specificity under MI

Moving on to the case of perfect MI, the various methods' specificity is shown in Figure 9. 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 9 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 7. In other words, susceptibility to increases in n is lower than in the presence of non-invariance. Ignoring the overall poorly performing J and CR methods, at least 80% of all items are correctly classified as being invariant. Here again, the BV method proves to be the most specific method while also remaining robust to changes of n. The remaining methods rank similarly compared to their overall ranking in terms of specificity. Yet, differences between methods are less pronounced in the setting of perfect MI.

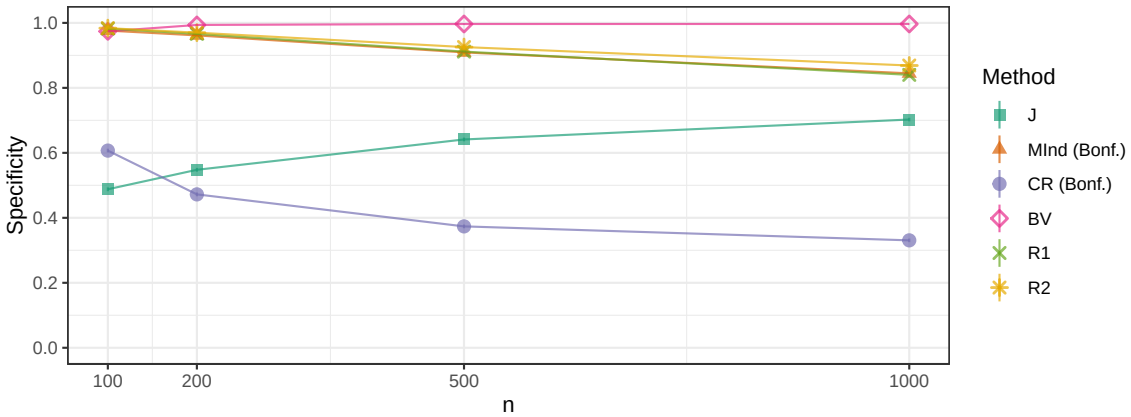

Figure 9: *Specificity of different detection methods as a function of n under perfect MI. Vertical lines represent 95% Clopper-Pearson confidence intervals.*

5.3.4 Summary and Discussion

The findings shown above have several implications for the use of these detection methods in the field. Broadly speaking, the methods can be put in three camps. First, the J and CR approaches don't work well in any setting and can thus not be recommended at all. They exhibit too many weaknesses, either theoretically or in terms of their performance in the simulation study. Second, the BV and MInd approaches perform decently in some settings but not in others. Particularly, the BV approach is a good choice when violations of MI affect both loadings and intercepts, but not when just one of the two is affected. On the other hand, the MInd approach works well for individual, but not for simultaneous violations. Either way, both are problematic because researchers generally don't know whether their items violate metric and/or scalar MI. They can thus not be recommended very generally. However, an argument can be made in favor of the BV approach in that it has consistently high specificity: It may not be able to detect a lot of non-invariant items, but it may still be sufficient to identify a starting point for model development. However, items that were not identified to be non-invariant should not be interpreted as being invariant and global tests should be used as a safeguard. The relatively poor specificity of the MInd approach, on the other hand, implies that it is unclear how to interpret its high number of positive classifications. Furthermore, although purely speculative, the case of simultaneous violations of MI may be practically more relevant than isolated violations. I thus consider the BV approach the best method among the existing methods as long as four or more items are used. Finally, the two versions of the R approach form the third camp. In terms of performance, it is clear that the R2 approach is superior to all other methods irrespective of the type of MI violation. It is also generally less affected by simulation parameters that worsen other methods' performance and should thus be considered as the preferred detection method. The R1 approach is in principle consistent across different types of MI violations, but is very susceptible to some of the other simulation parameters, in particular the group size. Despite the fact that there is no reason to opt for R1 over R2, it is put in the same group because it serves as the foundation for R2.

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 measurement models for populist attitudes in the CFA framework. 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 the application 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) note 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 despite their interest 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 MI of their models. 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 a selection of the models.

6.1 Synopsis of Castanho Silva et al. (2020)

Castanho Silva et al. (2020) compare seven existing measurement models of populist attitudes using original survey data from nine countries[19] in a CFA framework. In their comparison, the authors consider several properties relating to the models' internal coherence, cross-national validity, conceptual breadth, and external validity. Crucially, each of the 2556 respondents was subjected to every question that is required by any of the models so that the models can be fitted using a single sample, which makes their comparison more straightforward. However, due to some missing data, the number of observations per model is around 2150 with an average group size of around 250. 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. 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).[20] All of these studies build on Mudde's (2004) definition of populism as

[19] Brazil, France, Greece, Ireland, Italy, Mexico, Spain, United Kingdom, and the United States.

[20] An 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

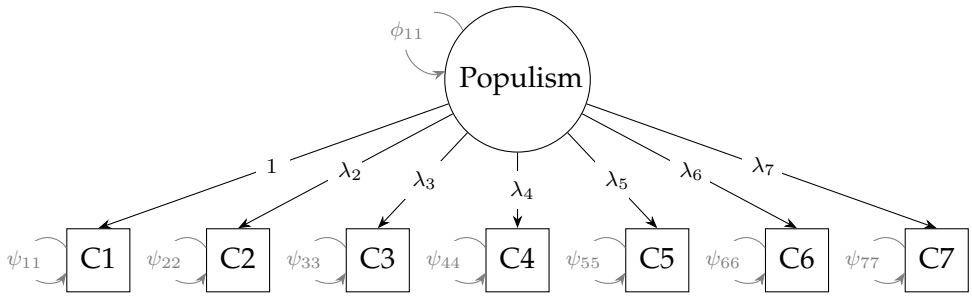

Figure 10: *Graphical depiction of Hobolt et al.'s (2016) populism model. Note that intercepts were omitted in this visualization.*

"a thin-centered 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 10 and 11, respectively, where intercepts are excluded in the visualizations, but not the models, for the sake of clarity.

Figure 10 illustrates Hobolt et al.'s (2016) model which is very straightforward: It incorporates a single latent variable, i.e. populism, and contains seven items $C1, ..., C7$ corresponding to seven survey questions. For each statement shown in Table 6, respondents are asked to indicate how much they agree on a scale from 1 (disagree) to 5 (agree). Furthermore, all items are assumed to load on the single latent factor and to have uncorrelated errors. Put differently, when fitting the model, each entry of $\mathbf{\Lambda}$ is estimated freely with the exception of a marker variable while the off-diagonal entries of $\mathbf{\Psi}$ 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.

On the other hand, Castanho Silva et al.'s (2018) model takes a very different approach. Their starting point is to argue that populism is a multidimensional concept, consisting of the latent concepts *anti-elitism* (ANT), *people-centrism* (PPL), and *anti-pluralism* (MAN). Figure 11 shows that they are modeled as pairwise correlated latent variables each with three items having pairwise uncorrelated errors. For the items, respondents were again asked to indicate their agreement with a set of statements which are listed in Table 7. Be-

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.*

sides 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 errors are modeled as pairwise uncorrelated, Castanho Silva et al. (2018) add an additional technical latent variable (Mtp) to account for some of the correlation for 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 $\mathbf{\Lambda}$ 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, $\mathbf{\Phi}$ 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 $\mathbf{\Psi}$ are constrained to zero as a result of the choice of pairwise uncorrelated specific variables. Although this step is not of particular interest for the purpose of this thesis, one might wonder how a single measure of populism can be obtained from this model where none of the latent variables is populism. Castanho Silva et al. (2018) propose to aggregate the estimates of the three main latent variables outside the CFA framework by rescaling them to the interval $[0, 1]$ and then taking their product. As a result, populism is seen as a multiplicative concept: High scores of populist attitudes require high scores on every underlying latent dimension (Castanho Silva et al., 2020). On the other hand, Hobolt et al.'s (2016) single-factor model represents an additive nature.

As part of their comparison of these and the remaining models, Castanho Silva et al. (2020) also assess the "cross-national validity", i.e. MI properties. 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.[21] The results of Castanho Silva et al.'s (2018) original

[21] One 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.

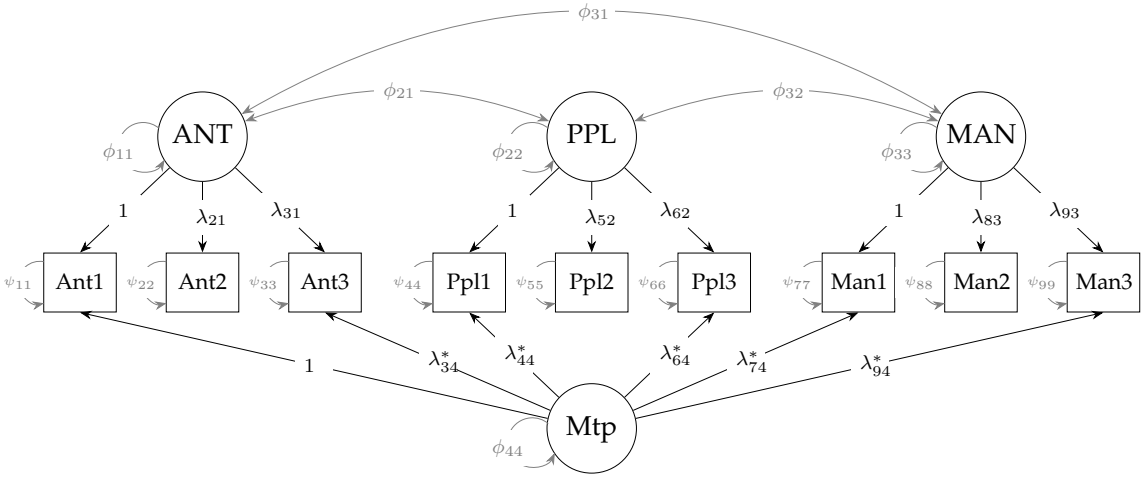

Figure 11: *Graphical depiction of Castanho Silva et al.'s (2018) populism model. Note that intercepts were omitted in this visualization. Further note that the five parameters relating to* Mtp *with an asterisk are constrained to equality.*

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.*

analysis are replicated in Table 8. In general, the results are not very encouraging: 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 problematic at best. 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 satisfies metric MI 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 cross-national validity.

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

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)*

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, scholars have introduced new survey questions when proposing new models; only in a few cases have existing survey questions been recycled (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 MI properties of each item. Applying the detection methods to their models gives them the possibility to determine which items contribute to 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 research with Castanho Silva et al.'s (2020) data could actually draw on the great number of readily available items from other models as replacements to improve a certain model. For a given model, problematic items may be replaced with theoretically comparable items from the entire catalogue of items across all models. In a next step, the modified model can then be analyzed with the detection methods again to verify if the replacement fares better in the given model. Conducting this procedure iteratively should finally yield a model that has better global MI properties. However, the other model properties that are considered by Castanho Silva et al. (2020) would also need to be analyzed in each iteration, which would go way beyond the scope of this thesis. I leave this to future research.

using real-world data.

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 the one by Castanho Silva et al. (2018), are multidimensional. Fortunately, generalizing the detection methods to the multi-factor case is straightforward: The existing 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 specific 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. For the R approaches, all that needs to be taken into account is that the residuals need to be obtained by regressing each item on all of its underlying latent variables together. The correlation component is then also tested for all of these latent variables in the regression.

However, there are a few discretionary choices that warrant mentioning: First, items that are modeled to relate to more than one latent variable are classified as being non-invariant if they are found to be non-invariant for any of their 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 under-identified comparison model was generated is arbitrarily classified as being invariant. This decision is somewhat justified by the fact that this is generally the null hypothesis of MI testing. Similarly, in the stepwise method R2, an item isn't removed if it is the last one remaining for a given latent variable. 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 constrained relationships between Mtp and its items are not tested for non-invariance 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 justification for disregarding these relationships 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.

6.3 Results

The results with regard to the MI properties of the items in Castanho Silva et al.'s (2018) and Hobolt et al.'s (2016) models generally indicate that several items are non-invariant. Given the fact that none of the models achieve global scalar MI and few metric MI, this

doesn't come as a surprise. In the following, I focus on the models by Hobolt et al. (2016) and Castanho Silva et al. (2018). 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 contain the detection results with the methods set to detect any type of MI violation. For the remaining five columns, the metric MI versions were used where only the invariance properties with respect to the loadings of each item are studied. With the exception of the BV method, all other methods apply a suitable Bonferroni or Bonferroni-Holm multiple-testing correction. Note that the J approach was omitted on the grounds of its theoretical weaknesses and poor performance in the simulation study.

		Metric & Scalar					Metric				
Item	**Survey**	**MInd**	**CR**	**BV**	**R1**	**R2**	**MInd**	**CR**	**BV**	**R1**	**R2**
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). The Bonferroni-adjusted versions of the MInd and CR approaches were used.*

Table 9 contains the results for the model by Hobolt et al. (2016). At a first glance, there is broad agreement between the different detection methods. Beginning with the detection of violations of metric and scalar violations, all methods unanimously classify items C3, C5, and C7 as non-invariant. Even more promising is the fact that the methods that achieved the highest specificity in the simulation study, i.e. the BV and R2 approaches, are completely unanimous in all their classifications. According to these two methods, item C1 can also be considered as non-invariant. At the same time, the MInd and CR approaches differ by classifying additional items as non-invariant, which is in line with the findings with regard to these methods in the simulation study. The main takeaway is that modelers should start with the investigation, replacement, modification, or removal of items C1, C3, C5, and C7 for further model development.

Turning to the final five columns, it is clear that the setting of the methods to the detection of violations of metric MI indeed results in fewer non-invariant classifications. This was expected theoretically and can be seen as a tentative indicator of these versions' validity. Yet, the results are not as straightforward as before: While the MInd, CR, and BV approaches exhibit some coherence, the R approaches classify no items as non-invariant. This is particularly puzzling because the overall model does not satisfy global metric MI. An intuitive explanation is that multiple items violate metric MI to a small degree such that at the model level, global MI is violated. Indeed, the p-values for items C1 and C5 in the R1 approach are very close to the rejection level. These two items were also identified by the BV method and should thus be investigated further for addressing metric non-invariance.

		Metric & Scalar					Metric				
Item	**Survey**	**MInd**	**CR**	**BV**	**R1**	**R2**	**MInd**	**CR**	**BV**	**R1**	**R2**
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). The Bonferroni-adjusted versions of the MInd and CR approaches were used.*

Similarly, Table 10 contains the results for Castanho Silva et al.'s (2018) model. Contrary to the findings above, there is considerable disagreement between the methods. Starting with metric and scalar invariance, most notable is the fact that the BV and R approaches are no longer in agreement. While the BV approach classifies a reasonable set of three items as non-invariant, the R1 approach comes to the conclusion that almost no item satisfies MI, and the R2 approach also classifies a large number of items as non-invariant. The only agreement between the well-performing methods BV and R2 is the item Ppl2, which may be investigated further. However, given the numerous disagreements, it seems prudent to refrain from further analysis of the metric and scalar invariance of the remaining items.

Turning to metric MI, things look slightly better: even the MInd approach with its low specificity classifies several items as being invariant. Further, the loading of item Ant2 is unanimously identified as non-invariant by all methods. Man2 is further chosen by both the BV and R2 method. However, this is where the agreement ends again. The starting points for further investigation and model development are thus Ant2 and Man2, but given the still considerable disagreement, the results should certainly be taken with a grain of salt. On a more general note, it is noteworthy that any items are classified as non-invariant by the metric versions despite the global null hypothesis of metric MI not being rejected for Castanho Silva et al.'s (2020) model. This is almost the flip side of the finding with regard to Hobolt et al.'s (2016) model which violates global metric MI while at the same time few items were classified as violating metric MI by the metric versions. Yet, this also shows that the detection methods may be beneficial and provide further insight by a more fine-grained look at the items' MI properties even when global MI conditions are met.

6.4 Summary and Discussion

The results have shown that - in principle - the detection methods can be applied to more complex CFA models than those used in the simulation study. For both models, the de-

tection 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 BV and R2 methods 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 or that many classifications indicate non-invariance. The results should therefore be viewed as tentative and as a first step. Yet, 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 accordingly.

7 Conclusion

The focus of this thesis has been the detection of items violating measurement invariance (MI) in CFA models. If gone undetected, violations of MI pose a serious threat to latent variable inference in a multi-group setting. Detecting non-invariant items during the model development phase enables researchers to account for this issue by revising, replacing, or removing these problematic items. Several detection methods have been proposed in the literature, but their performance has not been studied comprehensively. As a result, a lot of uncertainty remains as to which method should be preferred generally or in specific settings. Moreover, the existing methods require detailed knowledge of likelihood theory and the specification of numerous alternative models for model comparison. Both shortcomings of the existing methods pose a serious hurdle for researchers relying on CFA and may explain the persisting neglect of MI in applied research (c.f. Davidov et al., 2014).

Addressing these shortcomings of the literature and the existing methods, this thesis makes two key contributions. First, two versions of a novel approach for detecting non-invariant items (R1 & R2) have been introduced. Importantly, the novel approach is considerably more accessible and easier to implement than existing methods. Second, the novel approach and existing methods have been put to the test in a simulation study. The results provide researchers with guidance for the choice of a detection method. Furthermore, a secondary contribution of this thesis is a brief application of the detection methods to CFA models measuring populist attitudes using real-world data.

The accessibility of the R1 and R2 methods stems from their reliance on residuals of the linear relationships between latent variables and items in CFA models. In general, applied researchers are very familiar with the concepts of ANOVA and linear regression, which are the foundation of the detection methods. Furthermore, in the simple implementation R1, only a single model has to be fitted compared to the high number of models required for the existing methods. I show theoretically that violations of MI result in clearly distinguishable patterns in the CFA residuals and devise a statistical test of each item's MI properties. Finally, the possibility of visualization of the residual patterns is another advantage of this approach that may make it more accessible.

In the simulation study, data were generated to emulate different types of MI violations while varying several parameters such as the numbers of items, non-invariant items, and observations. The results show that most of the existing methods perform poorly. Of the four, only the BV approach has satisfyingly high sensitivity and specificity in some settings. Yet, it is still outperformed by the step-wise version of the novel approach (R2) which is more sensitive and only slightly less specific than the BV method. In addition, the R2 approach performs well for all types of MI violation, which is the biggest weakness of the BV approach. Moreover, in the case of perfect MI, all of the methods have few false positives. These findings have two implications for users of CFA models in multi-group settings. First, the question of which detection method should be used can be unequivocally an-

swered in favor of the R2 approach. In the case where high specificity is paramount, e.g. if false discoveries are particularly costly, and latent variables are modeled with more than three items, scholars can consider diverting to the BV approach, but should be aware of its potentially low sensitivity. Second, item-level detection methods of non-invariance should always be used during model development. In the best-case scenario of perfect MI, where no items suffer from non-invariance, the high specificity of all detection methods ensures that their use comes at very low cost.

The application of the detection methods to several CFA models measuring populist attitudes primarily shows that the detection methods can also be used for complex CFA models. In so doing, it also makes a contribution to the substantive study of populism by detecting items in existing models that are problematic in terms of their MI properties. Focusing on two specific measurement models, the detection methods have shown which items should be the starting point for model improvement by modification, replacement, removal, or at least further investigation. Notwithstanding, the application also shows how thorny the study of MI can be in the field: Considerable disagreement between the detection methods complicates the interpretation of their classifications. While some of these disagreements may be explained by the methods varying performance, even the two best methods, BV and R2, can disagree tremendously. However, given that the classifications in this application are only the first step of model improvement, this exercise is not completely useless and for several items, the detection methods are indeed consistent. Moreover, the application section also shows how more elaborate versions of the detection methods can be used to distinguish between the different types of MI violations.

The findings from both the simulation study and the application of the detection methods also open new avenues for further research. First, the simulation study may be extended vastly. To limit the scope of this thesis as well as the computational demand of the simulation, I focused on CFA models with a single latent variable and independent errors. Important insights may be gained from studying a broader range of CFA models that include more complex structural features. Second, and similarly, the simulation study may be extended to study the metric versions of the detection methods. While they certainly have the potential to provide useful information, their validity remains uncertain. Third, the detection methods themselves may be further refined to rank items by the degree of MI violation. This would aid the decision about which items should be the focus in the case when multiple items exhibit issues with MI. This would be relatively straightforward for some methods, for example by ranking the p-values of the item-specific hypothesis tests (R1, MInd) or by considering the order in which items are removed (R2). Finally, further substantive research may be conducted by salvaging the vast number of items collected by Castanho Silva et al. (2020) for model improvement. The battery of available items could be used as replacements for problematic items in one of the more promising 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/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. *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_FinalReport.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 Inglehart'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 & Opposition*, *39*(4), 541–563.

Mulaik, S. A. (1987). A brief history of the philosophical foundations of exploratory factor analysis. *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/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. *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/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. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Lawrence Erlbaum Associates, Inc.

Thomson, G. H. (1939). *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 $\boldsymbol{Y}$

$$\boldsymbol{Y} = \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon}. \tag{93}$$

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

$$\boldsymbol{\Sigma}(\theta) = \boldsymbol{\Lambda\Phi\Lambda}^\mathsf{T} + \boldsymbol{\Psi}, \tag{94}$$

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

Given a realization $\boldsymbol{y}_j$ of the p items for a single observation j, the likelihood can be written as

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

and the joint likelihood for n realizations as

$$\mathcal{L}(\theta \mid \boldsymbol{y}_{1:n}) = (2\pi)^{-np/2} \det(\boldsymbol{\Sigma}(\theta))^{-n/2} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}\boldsymbol{y}_j^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{y}_j\right).$$

Thus, the joint log-likelihood is given by

$$\begin{aligned}\ell(\theta \mid \boldsymbol{y}_{1:n}) &\coloneqq \log \mathcal{L}(\theta \mid \boldsymbol{y}_{1:n}) \\ &= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Sigma}(\theta)) - \frac{1}{2}\sum_{j=1}^{n}\boldsymbol{y}_j^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{y}_j.\end{aligned}$$

Next, the last term of the log-likelihood can be rewritten as

$$
\begin{aligned}
\frac{1}{2}\sum_{j=1}^{n}\boldsymbol{y}_j^\intercal\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{y}_j &= \frac{1}{2}\sum_{j=1}^{n}\operatorname{tr}\left(\boldsymbol{y}_j^\intercal\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{y}_j\right) \\
&= \frac{1}{2}\sum_{j=1}^{n}\operatorname{tr}\left(\boldsymbol{y}_j\boldsymbol{y}_j^\intercal\boldsymbol{\Sigma}^{-1}(\theta)\right) \\
&= \frac{n}{2}\operatorname{tr}\left(\sum_{j=1}^{n}n^{-1}\boldsymbol{y}_j\boldsymbol{y}_j^\intercal\boldsymbol{\Sigma}^{-1}(\theta)\right) \\
&= \frac{n}{2}\operatorname{tr}\Bigg(\underbrace{\sum_{j=1}^{n}n^{-1}\boldsymbol{y}_j\boldsymbol{y}_j^\intercal}_{=:\operatorname{diag}(\boldsymbol{S})}\boldsymbol{\Sigma}^{-1}(\theta)\Bigg) \\
&= \frac{n}{2}\operatorname{tr}\left(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}(\theta)\right)
\end{aligned}
$$

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 the diagonal of the sample covariance matrix $\boldsymbol{S}$ can be written as just $\boldsymbol{S}$ 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:

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

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

8.1.1 Connection with the Wishart Distribution

It can further be shown 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 $\boldsymbol{\Sigma}$. Let $\boldsymbol{V} \sim \mathcal{W}_p(\boldsymbol{\Sigma}, n)$ be a positive definite matrix with pdf

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

where $\Gamma_p(\cdot)$ is the multivariate gamma function. The likelihood of the observed sample covariance matrix $\boldsymbol{S}$ can thus be written as

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

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 with structure $\mathcal{M}$. Let $\mathcal{M}^{\text{base}}$ be the corresponding baseline model for which all items are modelled to be pairwise uncorrelated and without any underlying latent variables. In other words, $\mathcal{M}^{\text{base}}$ is a model with only item intercepts and item variances. Formally, the CFI can be defined as

$$\mathrm{CFI}\left(\mathcal{M}\right) \coloneqq 1 - \frac{\max\{T(\mathcal{M}), 0\}}{\max\{T(\mathcal{M}), T(\mathcal{M}^{\text{base}}), 0\}} \in [0, 1] \tag{95}$$

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

$$T(\cdot) \coloneqq nF(\mathcal{M}) - \mathrm{df}_{\mathcal{M}} \tag{96}$$

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 Derivation of Regression Coefficients γ and β

Recall that for a single item Y that is linearly regressed on the latent variable η, we defined a regression intercept γ and slope β. For the understanding of residual behavior on which the novel detection method is founded, it is instructive to see how these regression parameters relate to the group-specific intercepts τ^l and loadings λ^l of the data-generating process. In the following, we consider the most general case of MI violation, where both the intercepts and loadings may vary freely across groups. We thus have the group-specific DGP

$$Y^l = \tau^l + \lambda^l \eta^l + \varepsilon \tag{97}$$

for each group $l = 1, ..., g$. Suppose that each latent variable η^l has expectation $\mathbb{E}\left[\eta^l\right] = \mu^l$ and variance $\mathrm{Var}\left(\eta^l\right) = \phi^l$. Further suppose that group-membership is determined by a

multinomial random variable G, such that

$$G \overset{iid}{\sim} \text{Multinomial}(1, \boldsymbol{\pi}) \tag{98}$$

where

$$\boldsymbol{\pi} = (\pi^1, ..., \pi^g)^T = \frac{1}{N}(n^1, ..., n^g)^T \tag{99}$$

are membership probabilities or the relative frequencies of each group in our pooled or aggregated data.

We can thus define the aggregated Y as

$$Y \coloneqq \sum_{l=1}^{g} GY^l \tag{100}$$

and the aggregated η as

$$\eta \coloneqq \sum_{l=1}^{g} G\eta^l. \tag{101}$$

Note that the expectation of η is given by

$$\mu \coloneqq \mathbb{E}[\eta] = \sum_{l=1}^{g} \mathbb{E}\left[G^l\eta^l\right] = \sum_{l=1}^{g} \pi^l\mu^l \tag{102}$$

which is a weighted average of the group-specific latent means.

The pooled variance is given by

$$\begin{aligned}
\phi \coloneqq \text{Var}(\eta) &= \text{Var}\left(\sum_{l=1}^{g} G^l\eta^l\right) \\
&= \sum_{l=1}^{g} \pi^l \left(\phi^l + (1-\pi^l)(\mu^l)^2 - \sum_{\substack{k=1 \\ k\neq l}}^{g} \pi^k\mu^l\mu^k\right) \\
&= \sum_{l=1}^{g} \pi^l \left(\phi^l + \mu^l(\mu^l - \mu)\right)
\end{aligned} \tag{103}$$

which is a weighted average of the group-specific variances and an additional term depending on the group-specific latent means and their deviation from the pooled latent mean.

With these results, we can now derive the parameters of the regression of Y on η in the

aggregate. Beginning with the slope, we have that

$$\begin{aligned}
\beta &= \frac{\operatorname{Cov}(Y,\eta)}{\operatorname{Var}(\eta)} = \frac{1}{\phi}\operatorname{Cov}\left(\sum_{l=1}^{g} G^l Y^l, \sum_{k=1}^{g} G^k \eta^k\right) \\
&= \frac{1}{\phi}\sum_{l=1}^{g}\sum_{k=1}^{g} \operatorname{Cov}\left(G^l(\tau^l + \lambda^l\eta^l + \varepsilon), G^k\eta^k\right) \\
&= \frac{1}{\phi}\sum_{l=1}^{g}\left(\operatorname{Cov}\left(G^l(\tau^l + \lambda^l\eta^l + \varepsilon), G^l\eta^l\right) + \sum_{\substack{k=1\\k\neq l}}^{g} \operatorname{Cov}\left(G^l(\tau^l + \lambda^l\eta^l + \varepsilon), G^k\eta^k\right)\right) \\
&= \frac{1}{\phi}\sum_{l=1}^{g}\pi^l\Big(\tau^l\big(\mu^l - \mu\big) + \lambda^l\Big(\phi^l + \mu^l\big(\mu^l - \mu\big)\Big)\Big).
\end{aligned} \tag{104}$$

For the intercept, we have

$$\begin{aligned}
\gamma &= \mathbb{E}[Y] - \mathbb{E}[\beta\eta] \\
&= \mathbb{E}\left[\sum_{l=1}^{g} G^l Y^l\right] - \beta\mu \\
&= \sum_{l=1}^{g}\mathbb{E}\left[G^l\right]\mathbb{E}\left[Y^l\right] - \beta\mu \\
&= \sum_{l=1}^{g}\pi^l\left(\tau^l + \lambda^l\mu^l\right) - \beta\mu \\
&= \sum_{l=1}^{g}\pi^l\left(\tau^l\left(1 - \frac{\mu\big(\mu^l - \mu\big)}{\phi}\right) + \lambda^l\left(\mu^l - \frac{\mu\Big(\phi^l + \mu^l\big(\mu^l - \mu\big)\Big)}{\phi}\right)\right).
\end{aligned} \tag{105}$$

In essence, this shows that both the intercept and slope of the pooled regression are both a weighted average of the group-specific intercept and loading with complicated weights.

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

Below are the results of applying the detection methods to the remaining populism models in Castanho Silva et al.'s (2020) study. Note that for the measurement models by Oliver and Rahn (2016) and Stanley (2011) some convergence issues exist for some model structures that the CR approach creates, leading me to exclude these results.

	Metric & Scalar					Metric				
Survey	**MInd**	**CR**	**BV**	**R1**	**R2**	**MInd**	**CR**	**BV**	**R1**	**R2**
akker1	•			•	•		•			
akker2	•	•		•				•		
akker3		•						•		
akker4	•	•	•	•	•			•		
akker5	•	•	•	•	•			•		
akker6	•	•	•	•	•			•		

Table A1: *Items classified as non-invariant (•) in Akkerman et al. (2014). The Bonferroni-adjusted versions of the MInd and CR approaches were used.*

	Metric & Scalar					Metric				
Survey	**MInd**	**CR**	**BV**	**R1**	**R2**	**MInd**	**CR**	**BV**	**R1**	**R2**
es1	•		•							
es2	•	•	•	•	•			•		
es3		•			•					
es4	•	•	•	•	•			•		

Table A2: *Items classified as non-invariant (•) in Elchardus and Spruyt (2016). The Bonferroni-adjusted versions of the MInd and CR approaches were used.*

	Metric & Scalar					Metric				
Survey	**MInd**	**CR**	**BV**	**R1**	**R2**	**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). The Bonferroni-adjusted versions of the MInd and CR approaches were used. Note that CR results are excluded due to convergence issues.*

	Metric & Scalar					Metric				
Survey	**MInd**	**CR**	**BV**	**R1**	**R2**	**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). The Bonferroni-adjusted versions of the MInd and CR approaches were used.*

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

Table A5: *Items classified as non-invariant (•) in Stanley (2011). The Bonferroni-adjusted versions of the MInd and CR approaches were used. 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):

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

Verfasst von (in Druckschrift):
Bei Gruppenarbeiten sind die Namen aller Verfasserinnen und Verfasser erforderlich.

Name(n):	**Vorname(n):**
Rieger	Pit

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)**
Zürich, 08.02.2022	

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