On the Investigation of Measurement Invariance: New Developments and a Causal Framework for Future Research

Philipp Sterner

München, 2025

On the Investigation of Measurement Invariance: New Developments and a Causal Framework for Future Research

Inauguraldissertation

zur Erlangung des Doktorgrades der Philosophie

an der Ludwig-Maximilians-Universität München

vorgelegt von

Philipp Sterner

aus München

München, 2025

Erstgutachter: Prof. Dr. Markus Bühner

Zweitgutachter: Prof. Dr. David Goretzko

Tag der mündlichen Prüfung: 07.02.2025

"It is a fundamental criterion for a valid method of isolating primary abilities that the weights of the primary abilities for a test must remain invariant when it is moved from one test battery to another test battery. [...] This criterion assumes that the several test batteries are given to the same population. The primary abilities that define a test in one population should be identical with the primary abilities which define it in a second population."

— Louis L. Thurstone, *The Vectors of Mind*, p. 55, 1935

"Is this possible?"

— Harold Hotelling, *handwritten marginal comment*

Contents

List of Figures

List of Tables

1 Abstract

Measurement invariance (MI) means that the psychometric measurement models underlying a psychological questionnaire or test are equivalent across different groups or time points. MI is a prerequisite of meaningful between-group comparisons of measurements that the questionnaire produces. This thesis contains three manuscripts on methodological issues surrounding MI.

Study 1 introduces a new method to investigate MI, called Exploratory Factor Analysis Trees (EFA trees). EFA trees enable a data-driven assessment of MI among multiple continuous and/or categorical covariates (e.g., age, gender, education). To do so, they employ a score-based recursive partitioning algorithm. The result is a tree-like structure with so-called leaf nodes that contain partitions of the whole data set within which the measurement models are equivalent (i.e., for which MI holds). EFA trees are demonstrated by means of toy examples and investigated more thoroughly in an extensive Monte-Carlo simulation. The results indicate that EFA trees can reliably identify non-invariance under different conditions, for example, different sample sizes, group-size ratios, types of covariates, number of covariates, and types of violations of MI.

Study 2 provides an overview and comparison of EFA-based approaches to investigate MI that have been developed in recent years. The manuscript addresses multi-group EFA, mixture multi-group EFA, multi-group exploratory factor alignment, and EFA trees. For each method, the strengths and weaknesses as well as the assumptions underlying the method are detailed. Additionally, multi-group factor rotation is illustrated as a method to resolve the rotational indeterminacy of the EFA model. The application of all EFA-based methods, combined with multi-group factor rotation, is demonstrated on an empirical data set from moral psychology. To facilitate the application for applied researchers, template code in three different statstistical software programs (R, Mplus, Latent Gold) is made publicly available. In addition, a new R package EFAtree is developed to more easily implement EFA trees in the software R.

Study 3 moves the investigation of MI from a purely statistical to a more theoretical focus. A framework based on causal inference, more specifically directed acyclic graphs (DAGs), is presented. This framework allows to incorporate assumptions about causes

of non-invariance in the statistical modeling process. By explicitly depicting these assumptions in a DAG, researchers can make informed modeling decisions. Ultimately, this allows to view MI as a substantively interesting topic of research by itself, instead of a statistical assumption that licenses further analyses. By means of a simulated and an empirical example, the application of the framework is demonstrated.

In the general discussion, the applicability of the developments of the three manuscripts is critically assessed. Additionally, topics for future methodological research on MI are discussed, specifically tailored fit index cut-offs, effect size measures, and longitudinal MI. The thesis closes with a note on recent criticism against the necessity of MI for meaningful inference about latent variables.

2 Zusammenfassung

In der Psychologie interessieren wir uns häufig für die Ausprägung eines psychologischen Konstrukts bei Personen, zum Beispiel wie extravertiert oder intelligent eine Person ist. Dies gilt sowohl für die psychologische Forschung als auch für diagnostische Settings, zum Beispiel in der Einzelfalldiagnostik im Rahmen einer Psychotherapie. Während in der Forschung meistens Mittelwertunterschiede in psychologischen Konstrukten zwischen Gruppen von Interesse sind, hat die Diagnostik häufig zum Ziel, akurate Aussagen über Einzelpersonen zu treffen. Unabhängig vom spezifischen Setting ist eine zentrale Voraussetzung für zuverlässige Aussagen über Konstrukte, dass deren Ausprägung adäquat gemessen werden kann. Hierfür greift die Psychologie hauptsächlich auf Fragebögen zurück. Die Antworten, die eine Person auf die Fragen (Items) eines Fragebogens gibt, lassen mithilfe psychometrischer Messmodelle Rückschlüsse auf zugrundeliegende latente Variablen zu, die in diesen Messmodellen das zu messende Konstrukt repräsentieren.

Psychologische Fragebögen haben viele Kriterien, die Aussagen über ihre Qualität zulassen, sogenannte Gütekriterien. Ein solches Gütekriterium ist Messinvarianz (MI). Ein Fragebogen ist messinvariant, wenn er für alle Personen mit derselben Ausprägung auf einem Konstrukt dasselbe beobachtbare Ergebnis (also dieselben Anworten auf Fragebogenitems) produziert. Genauer formuliert liegt MI dann vor, wenn die den Fragebogenitems zugrundeliegenden Messmodelle für alle Personen äquivalent sind, unabhängig von möglichen Gruppenzugehörigkeiten (z.B. Alter, Geschlecht, Herkunft oder Bildung). Das vorliegen von MI ist ein fundamentales Kriterium, um Vergleichbarkeit psychologischer Messungen sicherzustellen. Wenn ein Fragebogen abhängig von Gruppenzugehörigkeiten unterschiedliche Ergebnisse bei identischer zugrundeliegender Konstruktausprägung hervorbringt, sind latente Mittelwertvergleiche und andere statistische Inferenzen verzerrt.

Die vorliegende Dissertation beinhaltet drei Manuskripte, die methodologische Themen zu MI behandeln. In der ersten Studie werden *Exploratory Factor Analysis trees* (EFA trees) vorgestellt, eine neue Methode zur Untersuchung von MI basierend auf explorativer Faktorenanalyse (EFA). Die zweite Studie gibt einen Überblick über neue methodische Entwicklungen im Bereich EFA-basierter MI-Untersuchung und vergleicht

diese neuen Methoden mittels eines empirischen Beispiels. In der dritten Studie wird ein Framework präsentiert, mit dessen Hilfe Ursachen von Verletzungen von MI (d.h. Non-Invarianz) grafisch veranschaulicht werden können. Dieses Framework basiert auf grafischen Methoden der kausalen Inferenz und ermöglicht es, die Annahmen der Methoden aus den ersten beiden Studien darzustellen und zu testen.

Studie 1 stellt EFA trees vor. EFA trees sind eine datengetriebene und explorative Methode, um MI in Bezug auf mehrere Kovariaten zu untersuchen. Die untersuchten Kovariaten können dabei sowohl kontinuierlich (z.B. Alter) als auch diskret (z.B. Bildung) sein. Darüber hinaus müssen keine zu untersuchenden Gruppenkonstellationen a priori festgelegt werden. Der den EFA trees zugrundeliegende Algorihmus ist das sogenannte *model-based recursive partitioning*. Dies ist ein Likelihood-basierter Algorithmus, der die Stabilität der Parameter eines Messmodells über alle Kovariaten hinweg testet. Wenn der Algorithmus Instabilität in den Parametern identifiziert (mittels *structural change tests*), dann teilt er den Gesamtdatensatz in Teildatensätze auf. Dieser Prozess wird in den Teildatensätzen wiederholt, bis die Parameter in diesen stabil sind oder andere Stoppkriterien erfüllt sind. Das Ergebnis eines EFA trees ist eine baum-ähnliche Struktur, in deren „Ästen“ Teildatensätze mit invarianten Messmodellen sind. Durch Betrachten der Kovariaten und ihren entsprechenden Ausprägungen, an denen der Datensatz geteilt wurde, lassen sich Rückschlüsse darauf ziehen, zwischen welchen Gruppen die Messmodelle non-invariant sind (z.B. Personen im Alter von 30 Jahren oder jünger, und Personen älter als 30 Jahre). Die Anwendung von EFA trees wird anhand einfacher simulierter Beispiele demonstriert. Sie werden außerdem im Rahmen einer umfassenden Monte-Carlo Simulation evaluiert. Die Ergebnisse dieser Simulation zeigen, dass EFA trees zuverlässig unter verschiedenen Bedingungen Non-Invarianz identifizieren.

In Studie 2 wird ein Überblick über in den letzten Jahren entwickelte Methoden zur Untersuchung von Messinvarianz gegeben. Diese Methoden haben gemeinsam, dass sie auf der EFA basieren, wodurch sie sich von bisherigen Methoden unterscheiden, die hauptsächlich auf der konfirmatorischen Faktorenanalyse (CFA) basieren. Durch die Verwendung der EFA als zugrundeliegendes Messmodell verringert sich das Risiko, dass Modellfehlspezifikationen auftreten. Die unter Umständen zu strengen Annahmen

bezüglich der Ladungen in Messmodellen einer CFA können das Ergebnise von MI-Untersuchungen verzerren. In der EFA gibt es diese Annahmen nicht, wodurch zusätzlich ermöglicht wird, auch Nebenladungen hinsichtlich ihrer Invarianz zwischen Gruppen zu untersuchen. Konkret werden die folgenden Methoden betrachtet und verglichen: *Multi-Gruppen EFA*, *Mixture Multi-Gruppen EFA*, *Multi-Gruppen Exploratory Factor Alignment* und EFA trees. Für jede Methode werden die Vor- und Nachteile sowie statistische Annahmen, die hinter den Modellen stehen, präsentiert. Zusätzlich wird die *Multi-Gruppen Factor Rotation* vorgestellt. Alle EFA-basierten MI-Methoden haben, wie die gewöhnliche EFA, das Problem der *rotational indeterminacy* (rotationale Unbestimmtheit). Dies bedeutet, dass die Faktorlösungen beliebig rotiert (sprich: transformiert) werden können, wodurch zwar die Passung des Modells an die Daten gleich bleibt, die Interpretation der Lösung aber stark beeinflusst werden kann. Multi-Gruppen Factor Rotation löst diese Unbestimmtheit, indem die Ladungsmatrizen in allen Gruppen so rotiert werden, dass sowohl die Einfachstruktur innerhalb der Gruppen als auch die Ähnlichkeit der Ladungsmatrizen zwischen den Gruppen maximiert wird. Dies stellt eine Vergleichbarkeit der Lösungen zwischen den Gruppen sicher.

Alle Methoden werden anhand eines empirischen Beispiels aus der Moralpsychologie demonstriert. Um die Anwendung für Forschende zu erleichtern, wird der Analysecode aus drei verschiedenen Softwareprogrammen (R, Mplus und Latent Gold) zur Verfügung gestellt. Darüber hinaus wurde im Rahmen von Studie 2 ein neues R-Paket (*EFAtree*) entwickelt, mit dessen Hilfe die EFA trees aus Studie 1 mit minimalem Aufwand angewendet werden können. Die Ergebnisse der empirischen Anwendung zeigen, dass EFA-basierte Methoden geeignet sind, um Ladungsinvarianz in den Daten zu untersuchen. Sie können als sinnvolle Ergänzung zu CFA-basierten Methoden gesehen werden, vor allem wenn sie diesen vorangestellt werden. Ein zentrales Ziel zukünftiger Forschung sollte es sein, die Fülle an verfügbaren Methoden (EFA- und CFA-basiert) in einem für die Anwendungsforschung geeigneten Prozess zu ordnen und zu vereinen.

Studie 3 legt den Fokus bei MI-Untersuchungen weg von einer rein statistischen hin zu einer theoretischeren Betrachtung. Es wird ein Framework basierend auf kausaler Inferenz präsentiert, mit dessen Hilfe Annahmen über die Ursachen von Non-Invarianz grafisch veranschaulicht werden können. Im Speziellen werden hierfür *Directed Acyclic*

Graphs (DAGs) verwendet. DAGs sind grafische Objekte, die kausale Zusammenhänge zwischen Zufallsvariablen darstellen. Die Zufallsvariablen werden mit gerichteten Pfeilen verbunden, wobei ein gerichteter Pfeil zwischen A und B anzeigt, dass A einen direkten kausalen Effekt auf B hat. Es wird veranschaulicht, wie die in der Literatur zu Faktorenanalysen üblichen Pfadmodelle in DAGs für Messmodelle übersetzt werden können und wie sich Non-Invarianz in DAGs darstellen lässt. Auf Basis dieser Darstellungen können dann informiertere Modellierungsentscheidungen getroffen werden, als wenn lediglich herkömmliche Methoden wie die Multi-Gruppen CFA heuristisch angewendet werden. Darüber hinaus ermöglicht die grafische Veranschaulichung der angenommenen Ursachen von Non-Invarianz eine tiefergehende Untersuchung dieser möglichen Gründe. Dies erlaubt es Anwendungsforschenden, MI als inhaltich interessantes Konzept zu betrachten und so etwas über das zugrundeliegende Konstrukt zu lernen. Dies ist ein großer Vorteil gegenüber der herkömmlichen, rein statstischen Betrachtungsweise, in der MI lediglich als zusätzlich zu überprüfende Annahme vor den eigentlichen statistichen Analysen gesehen wird. Viel mehr enthält das Fehlen von Messinvarianz zwischen Gruppen Informationen darüber, wie diese Gruppen ein Konstrukt interpretieren oder wie es sich in diesen Gruppen im Erleben und Verhalten manifestiert.

Die Dissertation schließt mit einer allgemeinen Diskussion. In dieser wird die Anwendbarkeit der vorgestellten Methoden und des Frameworks aus Studie 3 kritisch betrachtet. Weitere methodologische Forschung, die vor allem konkrete Anleitungen zur Anwendung der Methoden sowie Softwareimplementierungen beinhaltet, ist notwendig, damit die Anwendungsforschung vollumfänglich von den Ergebnisse der drei Manuskripte profitieren kann. Zusätzliche inhaltiche Themen zukünftiger Forschung sind vor allem dynamische Grenzwerte von Fit-Indizes zur Erkennung von Non-Invarianz, Effektstärkemaße sowie die Konzeptualisierung von MI bei Betrachtung longitudinaler Daten. Abschließend wird auf kürzlich veröffentlichte Kritik an der Notwendigkeit von MI für die psychologische Forschung eingegangen.

3 General Introduction

In psychology, we are almost always interested in some psychological construct in humans, for example how extraverted or intelligent a person is. This pertains to both psychological research and diagnostic settings, like psychotherapeutic assessments. In research, we are often concerned with mean differences between groups on the constructs under investigation and in diagnostics, our goal is to make accurate statements about a single person. Regardless of the setting, a central prerequisite of unbiased statements about constructs is that they are adequately measured. To measure the value of a construct, psychologists commonly rely on questionnaires. The responses to questionnaire items (also: observed variables) are related to a latent variable that represents the construct we intend to measure (Lord & Novick, 1968). Because of their prominent role and psychology's dependence on them, it is crucial that the questionnaires we use to measure constructs are thoughtfully constructed (Wijsen et al., 2022). There are many definitions of a "good" measurement produced by a questionnaire, for example that they are objective, reliable, and valid (Bühner, 2021; Kline, 2015). Other aspects include feasibility, time efficiency, and fairness. Focusing more on the specific measurement properties of a questionnaire, one important part is that its measurements are invariant. *Measurement invariance* (MI) means that the questionnaire measures the same construct in the same way across all possible groups or time points (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Put differently, if two persons share the same *true score* on a specific construct, for example extraversion, an invariant extraversion questionnaire should produce the same *measured score* (in the form of item responses), regardless of other differences between these two persons, like gender, age, or education. The same applies to subsequent measurement occasions: if a construct is measured twice on the same individual and this individual's true score on the construct does not change over time, then the measured score must remain the same as well.

Consider a depression questionnaire that consistently produces higher measured scores for younger patients compared to older patients, even though they have the same level of depression severity (i.e., the same true score). In research, where often group differences in latent means between groups are analyzed, this lack of MI would lead to biased inference between younger and older patient groups (Putnick & Bornstein, 2016).

In settings of single-case assessments, like psychotherapy, younger patients would be overdiagnosed and older patients, in turn, underdiagnosed with depression, leading to an inappropriate or inequitable allocation of available treatment. Thus, it is crucial to establish MI so that our measurements allow for both meaningful comparisons in research and accurate assessments in diagnostic settings. Only if MI holds, we can be sure that any difference we observe between groups occurs due to true differences and not due to differences in measurement (Meuleman et al., 2023).

3.1 Manuscripts of this Thesis

This thesis comprises the following three manuscripts on the topic of MI:

1. Sterner, P., & Goretzko, D. (2023). Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 871–886. https://doi.org/10.1080/10705511.2023.2188573

2. Sterner, P., De Roover, K., & Goretzko, D. (2024). New Developments in Measurement Invariance Testing: An Overview and Comparison of EFA-based Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(1), 117–135. https://doi.org/10.1080/10705511.2024.2393647

3. Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A Causal Framework for the Comparability of Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 747–758. https://doi.org/10.1080/10705511.2024.2339396

All manuscripts were conceptualized and written by the first author of this thesis. David Goretzko was the supervising author in all three studies. Florian Pargent, Kim De Roover, and Dominik Deffner acted as co-authors. Their individual contributions are stated in all three manuscripts. Because all manuscripts are a team effort, the pronoun "we" will be used throughout this thesis and its manuscripts. If the pronoun "I" is used in the General Introduction or General Discussion, it is to indicate that the statement is the opinion of the author of this thesis.

The first study (Study 1) introduces a new approach called *exploratory factor*

analysis trees (EFA trees). EFA trees enable a data-driven assessment of MI among multiple continuous and/or categorical covariates (e.g., age, gender, education). To do so, they draw on a score-based recursive partitioning algorithm (Hothorn et al., 2006; Zeileis et al., 2008). The second study (Study 2) provides an overview and comparison of newly developed EFA-based approaches to investigate MI. Up until recently, MI was primarily assessed based on *confirmatory factor analysis* (CFA). By basing the investigation on EFA instead, specific facets of MI can be assessed in more detail without the assumption of a potentially too strict measurement model (Nájera et al., 2023). Finally, the third study (Study 3) details a framework based on *causal inference* that enables researchers to reason about the causes of a lack of MI (i.e., non-invariance). Only by making assumptions about underlying causal structures and possible causes of violations of MI we can see non-invariance as an interesting research finding by itself. To make full use of the methods presented in Studies 1 and 2, reasoning about the underlying structures as suggested in Study 3 is crucial to make informed modeling decisions.

The remainder of this thesis is structured as follows: First, we provide a more technical account of MI and its implications. Second, we present the current state of the literature on MI as well as the gaps which the three studies aim to fill. Third, the three studies are printed in full length. Forth and last, we discuss the results more generally by suggesting possible directions for future research and by addressing recent criticism against the necessity of MI for psychological science.

3.2 Factor Analysis

MI is defined as the equivalence of measurement models across groups or time points. Thus, to define MI more formally, we first have to introduce the measurement models underlying its investigations. MI is usually assessed based on multi-group exploratory or confirmatory factor analysis (MG-EFA and MG-CFA, respectively). Both models are instances of the more general multi-group common factor model (Jöreskog, 1971; Sörbom, 1974). EFA is used to freely uncover relations between manifest (i.e., observed) and latent (i.e., unobserved) variables (Goretzko et al., 2021) and CFA is used to test an assumed relation between these two types of variables. Let $\boldsymbol{x}_{i_g}$ be a p-dimensional random vector of observed variables (e.g., responses to

questionnaire items) of observation i in group g (with $i \in \{1, ..., n_g\}$ and $g \in \{1, ..., G\}$). This vector can be seen as a linear combination of m latent variables (also: factors; e.g., extraversion or intelligence), that is,

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

Here, $\boldsymbol{\tau}_g$ is a p-dimensional vector of group-specific item intercept, $\boldsymbol{\Lambda}_g$ is a $p \times m$-dimensional matrix of group-specific factor loadings, $\boldsymbol{\xi}_{i_g}$ is a m-dimensional vector of latent factor scores, and $\boldsymbol{\epsilon}_{i_g}$ is a p-dimensional vector of error terms. $\boldsymbol{\tau}_g$ describes the item means when the latent variables are equal to 0 and $\boldsymbol{\Lambda}_g$ quantifies the strength of the linear relation between the manifest and latent variables. Usually, distributional assumptions are made for estimation purposes (Jöreskog, 1967): The latent factor scores are assumed to be multivariate-normally distributed, that is, $\boldsymbol{\xi}_{i_g} \sim MVN(\boldsymbol{\alpha}_g, \boldsymbol{\Phi}_g)$, where $\boldsymbol{\alpha}_g$ and $\boldsymbol{\Phi}_g$ denote the group-specific factor means and (co-)variances, respectively. The error terms are also assumed to be multivariate-normal, that is, $\boldsymbol{\epsilon}_{i_g} \sim MVN(0, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\Psi}_g$ is a $p \times p$ matrix which contains the group-specific unique variances of the observed variables on its diagonal and the value 0 on all off-diagonal entries (i.e., no correlated errors are allowed). Error terms and factor scores are assumed to be independent ($\mathbb{E}(\boldsymbol{\xi}\boldsymbol{\epsilon}^\top) = \mathbf{0}$). The group-specific model-implied covariance matrix is then defined as $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Phi}_g \boldsymbol{\Lambda}_g{}^\top + \boldsymbol{\Psi}_g$; that is, variation in item responses can be decomposed into a part explained by the latent variable and into a unique error part.

When EFA is used, there are no zero-constraints in $\boldsymbol{\Lambda}_g$ and the estimated loading matrix is only determined up to admissible rotations. That is, infinitely many solutions exist which all have the same fit to the data but lead to different interpretations of the solution. This rotational indeterminacy has to be resolved per group by specifying a rotation criterion (for an overview of rotation criteria, see Browne, 2001). In single-group settings, usually a rotation criterion aiming at an independent clusters structure ("simple structure"), like geomin, is used. That is, the solution is rotated such that all observed variables show a high loading on one factor and near-zero loadings on all other factors. In multi-group settings, the agreement of solutions *between groups* has to be considered in addition to simple structure *per group* in order to meaningfully compare

solutions across groups (De Roover & Vermunt, 2019). In the studies in this thesis, two different approaches to tackle the problem of rotational indeterminacy in multi-group settings are used: in study 1, we use *elastic net regularization* (Zou & Hastie, 2005), and in study 2, we use *multi-group factor rotation* (De Roover & Vermunt, 2019).

3.3 A Formal Definition of Measurement Invariance

MI is formally defined as the equivalence of measurement models across any group defined by a covariate V, that is

$$f(\boldsymbol{x}_{i_g}|\boldsymbol{\xi}_{i_g}, V_i) = f(\boldsymbol{x}_{i_g}|\boldsymbol{\xi}_{i_g}). \tag{2}$$

where $f(\cdot)$ is the probability density function (Mellenbergh, 1989; Meredith, 1993). Given the latent variable(s) $\boldsymbol{\xi}_{i_g}$, the item responses $\boldsymbol{x}_{i_g}$ of person i in group g are independent of this person's group membership V_i (i.e., their age, gender, education,...). Because the latent variable $\boldsymbol{\xi}_{i_g}$ is unobservable, this conditional independence cannot be tested directly (in Study 3, we provide a definition of MI from the perspective of *causal inference*). Instead, to test MI, researchers fit a series of increasingly constrained models (in the form of Equation 1) to test various nested levels of MI (Van de Schoot et al., 2012). The idea is that if the parameters of a measurement model are equivalent across groups (i.e., if MI holds), fitting the model in all groups with equivalence constraints on some parameters (e.g., equal loadings across groups), should not decrease the combined model fit. If the fit, however, drops when adding equality constraints between groups, this is an indication that some parameters are not equivalent across groups and should be allowed to have group-specific values.

3.3.1 The Different Levels of Measurement Invariance

Usually, four nested levels of MI are consecutively considered (Meredith, 1993; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000).

3.3.1.1 Configural invariance.

Configural invariance, also called invariance of model form, entails that the same pattern of free and fixed-to-zero loadings holds across groups. That is, the positions of the entries in $\boldsymbol{\Lambda}_g$ that are estimated freely and the entries that are fixed to zero

should be equivalent across groups. Additionally, the number of latent variables (i.e., the dimension of $\boldsymbol{\xi}_{i_g}$) should be equal in each group. It is worth mentioning that this level of invariance only fully applies to investigations of MI in the context of CFA, where we impose constraints on $\boldsymbol{\Lambda}_{\boldsymbol{g}}$. In EFA, where all entries of $\boldsymbol{\Lambda}_{\boldsymbol{g}}$ (i.e., all main- and cross-loadings) are freely estimated, there is no pattern of free and fixed loadings. A more relaxed definition might be that the position of loadings estimated to be (exactly) zero and non-zero must be the same across groups. But because loadings will hardly ever be exactly zero and a loading of exactly zero can also be interpreted as the absence of cross-loadings, this definition could also be considered a form of metric invariance, which will be discussed in the following. Since EFA-based assessment of MI is still in its infancy, a formal definition of configural invariance for this type of methods has yet to be agreed on.

3.3.1.2 Metric invariance.

Metric invariance, also called weak invariance, means that the loadings are equivalent across groups (i.e., $\boldsymbol{\Lambda}_{\boldsymbol{g}} = \boldsymbol{\Lambda}_{\boldsymbol{k}} \; \forall \; g, k \in \{1, ..., G\}$ with $g \neq k$). Equal loadings across groups entail that the (linear) relations between observed and latent variables are equal across groups. This means that each observed variable or item contributes to the measurement of the latent variable by the same amount and that they share the same scale. If metric invariance is supported, latent relations (e.g., covariances between latent variables) can be compared across groups and one can proceed to test the next level of MI.

3.3.1.3 Scalar invariance.

Scalar invariance, also called strong invariance, means that, in addition to the loadings, the intercepts are equivalent across groups (i.e., $\boldsymbol{\tau}_{\boldsymbol{g}} = \boldsymbol{\tau}_{\boldsymbol{k}} \; \forall \; g, k \in \{1, ..., G\}$ with $g \neq k$). If scalar invariance holds, the items have the same origin (on the same scale if metric invariance holds as well) because the intercepts are equal to the item means if the value on the latent variable is zero. Scalar invariance is required for a meaningful comparision of latent means.

3.3.1.4 Residual invariance.

Residual invariance, also called strict invariance, is the last level of MI and means that, in addition to all equivalences above, the unique variances of the observed variables

(i.e., residual variances) are equal across groups (i.e., $\boldsymbol{\Psi_g} = \boldsymbol{\Psi_k}\ \forall\ g, k \in \{1, ..., G\}$ with $g \neq k$). Because the residuals are not part of the latent factor scores, this step is not a prerequisite of meaningful latent mean comparisons. For this reason, it is often not considered in investigations of MI. However, for some research questions and especially single-case assessments it might be interesting to assess because residual invariance entails that the latent variable is measured with the same precision across groups (i.e., equal item reliabilities across groups). As this thesis focuses on latent mean comparisons, residual invariance is not frequently addressed.

3.3.2 When is Measurement Invariance Supported?

The four models (configural, metric, scalar, residual) above are estimated consecutively, adding equality constraints on the parameters between groups in a stepwise manner. For each model, a combined measure of fit across all groups is calculated. If the fit of the model does not decrease substantially when adding equality constraints (e.g., equal loadings across groups), this is an indication of MI (Van de Schoot et al., 2012). Most often, the fit indices that are reported are the *root mean squared error of approximation* (RMSEA) and the *comparative fit index* (CFI). There are rough rules-of-thumb for changes in these fit indices that would indicate non-invariance (Chen, 2007; Cheung & Rensvold, 2002); for example, an increase of the RMSEA by at least 0.01 and/or a decrease of the CFI by at least 0.01 are an indication that MI is not supported. Rutkowski and Svetina (2014) proposed more liberal cut-offs in settings where the number of groups is high (i.e., above 10), especially for the investigation of metric MI (i.e., when adding equality constraints on the loadings); specifically, an increase of the RMSEA by at least 0.03 and/or a decrease of the CFI by at least 0.02. Cao and Liang (2022a) provided more details when testing MI in models with cross-loadings, evaluating the behavior of these fit indices under various model and data conditions (i.e., in EFA-based analyses). In addition to the inspection of fit indices, an exact model comparison by means of χ^2-difference tests (likelihood ratio test) is possible, too, because the models are nested. While this test is especially suitable to detect non-invariance in the presence of cross-loadings (Cao & Liang, 2022a), it might be too sensitive in large samples and detect practically irrelevant degrees of non-invariance (De Roover et al., 2022).

In general, while these cut-offs are helpful in reaching a binary decision regarding (non-)invarince, they should never be taken at face value. As with model evaluation in single-group settings, adequate cut-offs depend on sample size, model complexity, and other nuisance parameters like the absolute size of the loadings (Cao & Liang, 2022b; Goretzko et al., 2023; Partsch et al., 2024). Instead of simply adopting these fixed cut-offs, researchers should thus rather take into account as much information regarding MI as possible, discuss inconsistent results, and —most importantly— evaluate their results against the background of (model) plausibility and theoretical adequacy.

3.4 Current State of Measurement Invariance in Psychological Research

First, up until recently, MI has been primarily assessed based on CFA (for an overview of CFA-based methods, see Kim et al., 2017). While this allows to test assumed relations between observed and latent variables and their invariance across groups, it also limits the capabilities of investigations of MI with respect to metric MI. Because in CFA, cross-loadings in $\boldsymbol{\Lambda_g}$ are usually not considered, non-invariance due to cross-loadings cannot be assessed (De Roover et al., 2022). Additionally, the same, potentially too strict, measurement model has to be assumed across groups which can lead to model misspecifications (Nájera et al., 2023). To counteract these issues of CFA-based MI investigations, more methods based on EFA have recently been developed. In EFA, no strict measurement model has to be assumed, and thus, cross-loadings can be evaluated for invariance, too. This allows researchers to modify their models without having to repeatedly test altered versions of a CFA in a data-driven way.

Second, although numerous methods (e.g., Kim et al., 2017) and guides (e.g., Van de Schoot et al., 2012) to investigate MI are available, it is very rarely done in empirical studies (Maassen et al., 2023). This is problematic because it raises the question of how many latent mean differences reported in the literature occurred due to true differences and how many occurred due to differences in measurements. As a consequence, an important aspect of methodological research on the topic of MI should be to increase the prevalence of investigations of MI in psychological science.

The manuscripts of this thesis aim to extend recent advances and address current issues in research on MI. They introduce a new EFA-based method (Study 1), summarize

and compare EFA-based methods (Study 2), and propose a framework that assists in the meaningful application of both CFA- and EFA-based approaches (Study 3). In a sense, this thesis first expands the methodological toolbox of MI, before cataloging parts of this toolbox, and ultimately provides a potential guide on how to purposefully apply its tools.

3.4.1 Contribution of Study 1

The use of EFA in investigations of MI allows researchers to consider MI already in the earliest stages of questionnaire development. In this stage, EFA is used to uncover relations between observed and latent variables (Goretzko et al., 2021). This is a major advantage because at this stage, changes to the item pool or in item wordings are still possible. If questionnaires are developed as invariant as possible, issues with non-invariance in subsequent studies where they are applied might be prevented. Ideally, MI is considered with regard to many different covariates during questionnaire development. This ensures that it can be applied to compare measurements in a variety of contexts, regardless of specific group constellations (i.e., across different combinations of ages, genders, educational backgrounds, but also across regions and countries in cross-cultural settings). To enable fully exploratory investigations of MI with many covariates, in Study 1 we introduce EFA trees. EFA trees simultaneously evaluate multiple continuous (e.g., age) and categorical (e.g., education) covariates for MI in a data-driven manner. The underlying algorithm is so called model-based recursive partitioning, a method that repeatedly splits the data in order to increase the fit of a model estimated on the (partitioned) data (Hothorn et al., 2006; Zeileis et al., 2008). If EFA trees identify non-invariance of measurement models between groups, they split the data on the covariate which best explains this non-invariance. In doing so, they reveal non-invariant groups without the need to specify the groups that are to be compared in advance. This allows to investigate MI more broadly, both in terms of different group constellations and with regard to main- and cross-loadings, making them particulary suitable for questionnaire development. EFA trees join prior work in which model-based recursive partitioning has been combined with psychometric models. Most closely related, Brandmaier et al. (2013) introduced structural equation model trees (SEM trees), and others developed tree-based methods in an item response theory

paradigm (e.g., Komboz et al., 2018; Strobl et al., 2015).

3.4.2 Contribution of Study 2

MI is a prerequisite for meaningful latent mean comparisons and should therefore be routinely investigated in psychological studies. Unfortunately, the opposite is the case: it is very rarely considered (Maassen et al., 2023). Possible reasons for this lack of investigations of MI are surely diverse. However, one reason that could be ruled out is that there are not enough methods to properly assess MI. Beyond MG-EFA and MG-CFA mentioned above, there is a variety of methods that could be applied to investigate MI, each with specific strengths and weaknesses (e.g., Asparouhov & Muthén, 2014, 2023; De Roover et al., 2022; Kim et al., 2017). The choice of the optimal method depends on the data conditions and research goals at hand. This, in turn, might deter or confuse applied researchers, simply because it is difficult to maintain an overview over all available options; let alone choose the most appropriate method for an analysis. Kim et al. (2017) provided a comprehensive overview of CFA-based methods, facilitating the oversight of different options. Our contribution to the literature on MI in Study 2 is an overview and comparison of new developments in the EFA-based assessment of MI. By detailing the strengths and weaknesses of these recently developed methods, we aim at lowering the barriers to understand and apply them in research. To facilitate the application of these methods, we provide openly available *R*, *Mplus*, and *Latent Gold* code that researchers can use as a blueprint for their own analyses. Building on Study 1, we also developed an openly available R package, called *EFAtree* (available at https://github.com/philippsterner/EFAtree). The *EFAtree* package contains wrapper functions to grow EFA trees with minimal coding effort and helper functions to explore the results of EFA trees.

3.4.3 Contribution of Study 3

MI is often described as “essential” (Maassen et al., 2023) or a “prerequisite” for meaningful comparisons of latent means (e.g., De Roover, 2021; De Roover et al., 2022; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). While this is true, it hints at the current role of MI in psychological research: a statistical assumption that has to be investigated (more strictly: tested) *before* the actual analysis of interest. This neglects

the fact that (a lack of) MI can be an interesting research finding in itself. Discovering that the same questionnaire works differently across groups can provide information on how these groups interpret a construct or how it manifests itself in these groups (Fischer & Rudnev, 2024; Putnick & Bornstein, 2016). This notion is not new in the literature. For example, Putnick and Bornstein (2016) encouraged readers to not view investigations of MI as a "gateway test" (p. 87) and Maassen et al. (2023) cautioned to not take non-invariance as a "roadblock to further analysis" (p. 12). Both groups of authors argued that assessing MI can yield substantively relevant information to better understand differences between groups.

What is currently missing, however, is a framework that allows researchers to view MI as an interesting finding by itself. Without being able to reason about potential causes underlying a lack of MI, it is difficult to relate the results of investigations of MI back to the substantive analysis. Therefore, in Study 3, we propose a framework based on *causal inference*, more specifically *directed acyclic graphs* (DAGs), which allows researchers to explicitly depict their assumptions about potential violations of MI. This framework builds on recent work by Deffner et al. (2022) who presented a similar framework but for observed data (i.e., manifest variables). We extend their framework to typical psychological research settings. In psychology, we usually want to make claims on the construct level and, consequently, MI becomes an important aspect of the modeling process.

DAGs are graphical objects that allow us to visualize the causal relations between variables (Elwert, 2013; Pearl, 1988, 1998, 2012). We demonstrate how commonly used path diagrams from the linear SEM literature can be translated into DAGs and how non-invariance can be depicted by selection diagrams (Pearl & Bareinboim, 2014). Following graphical rules of DAGs, informed modeling choices can then be made to reason about and investigate potential causes of non-invariance. In this, the full potential of the methods introduced and presented in Studies 1 and 2 can be leveraged. Only by making informed modeling decisions based on explicitly stated assumptions, the full potential of a statistical method can be used (for an example of how to take into account the underlying assumptions of advanced methods, see Luong & Flake, 2023).

Similar to providing an overview of advanced methods in Study 2, we aim at increasing the prevalence of investigations of MI in psychological science. The proposed framework allows researchers to view a lack of MI as an interesting research finding, instead of just an additional test to license further analyses. This hopefully motivates researchers to think more thoroughly about potential causes of non-invariance and to explicitly state the assumptions underlying latent mean comparisons. In the following three chapters, Studies 1, 2, and 3 are printed in full length.

3.5 References

Asparouhov, T., & Muthén, B. O. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Asparouhov, T., & Muthén, B. O. (2023). Multiple Group Alignment for Exploratory and Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(2), 169–191. https://doi.org/10.1080/10705511.2022.2127100

Borsboom, D., Romeijn, J.-W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods*, *13*(2), 75–98. https://doi.org/10.1037/1082-989X.13.2.75

Borsboom, D., & Wijsen, L. D. (2017). Psychology's atomic bomb. *Assessment in Education: Principles, Policy & Practice*, *24*(3), 440–446. https://doi.org/10.1080/0969594X.2017.1333084

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion*. Pearson Deutschland.

Cao, C., & Liang, X. (2022a). Sensitivity of Fit Measures to Lack of Measurement Invariance in Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 248–258. https://doi.org/10.1080/10705511.2021.1975287

Cao, C., & Liang, X. (2022b). The Impact of Model Size on the Sensitivity of Fit Measures in Measurement Invariance Testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 744–754. https://doi.org/10.1080/10705511.2022.2056893

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

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

De Roover, K. (2021). Finding Clusters of Groups with Measurement Invariance: Unraveling Intercept Non-Invariance with Mixture Multigroup Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., & Vermunt, J. K. (2019). On the Exploratory Road to Unraveling Factor Loading Non-invariance: A New Multigroup Rotation Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 905–923. https://doi.org/10.1080/10705511.2019.1590778

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

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A Causal Framework for Cross-Cultural Generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3). https://doi.org/10.1177/25152459221106366

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_13

Fischer, R., & Rudnev, M. (2024). From MIsgivings to MIse-en-scène: the role of invariance in personality science. *European Journal of Personality*

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521.

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating Model Fit of Measure-

ment Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482.

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

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kline, P. (2015). *A Handbook of Test Construction (Psychology Revivals): Introduction to Psychometric Design*. Routledge. https://doi.org/10.4324/9781315695990

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166. https://doi.org/10.1177/0013164416664394

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. IAP.

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)

90002-5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2023). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research*, *52*(3), 1401–1419. https://doi.org/10.1177/00491241221091755

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

Partsch, M., Sterner, P., & Goretzko, D. (2024). *A Simulation Study on the Interaction Effects of Underfactoring and Nuisance Parameters on Model Fit Indices*. OSF. https://doi.org/10.31234/osf.io/qy2e3

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods & Research*, *27*(2), 226–284. https://doi.org/10.1177/0049124198027002004

Pearl, J. (2012). *The Causal Foundations of Structural Equation Modeling:* Defense Technical Information Center. https://doi.org/10.21236/ADA557445

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, *29*(4). https://doi.org/10.1214/14-STS486

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

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psy-*

chological Measurement, *74*, 31–57. https://doi.org/10.1177/0013164413498257

Sörbom, D. (1974). A General Method for Studying Differences in Factor Means and Factor Structure Between Groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Wijsen, L. D., Borsboom, D., & Alexandrova, A. (2022). Values in Psychometrics. *Perspectives on Psychological Science*, *17*(3), 788–804. https://doi.org/10.1177/17456916211014183

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

4 Study 1: Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates

Sterner, P., & Goretzko, D. (2023). Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 871–886. https://doi.org/10.1080/10705511.2023.2188573

The authors made the following contributions. Philipp Sterner: Conceptualization, Formal Analysis, Methodology, Visualization, Writing - Original Draft; David Goretzko: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

4.1 Abstract

Measurement invariance (MI) describes the equivalence of a construct across groups. To be able to meaningfully compare latent factor means between groups, it is crucial to establish MI. Although methods exist that test for MI, these methods do not perform well when many groups have to be compared or when there are no hypotheses about them. We suggest a method called *Exploratory Factor Analysis Trees* (EFA trees) that are an extension to *SEM trees*. EFA trees combine EFA with a recursive partitioning algorithm that can uncover non-invariant subgroups in a data-driven manner. An EFA is estimated and then tested for parameter instability on multiple covariates (e.g., age, education, etc.) by a decision tree based method. Our goal is to provide a method with which MI can be addressed in the earliest stages of questionnaire development or prior to analyses between groups. We show how EFA trees can be implemented in the software *R* using *lavaan* and *partykit*. In a simulation, we demonstrate the ability of EFA trees to detect a lack of MI under various conditions. Our online material contains a template script that can be used to apply EFA trees on one's own questionnaire data. Limitations and future research ideas are discussed.

4.2 Introduction

In psychometrics, measurement invariance (MI) describes the equivalence of measurements of a construct across groups (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). This concerns different groups of a population (e.g., women and men) or subsequent measurement occasions of the same group (e.g., pre- and post-treatment). If MI does not hold between two or more groups, it cannot be readily assumed that the construct of interest has the same meaning to people between these groups. Consequently, analyses like comparisons of means and variances across groups or measurement occasions will not be meaningful or will even yield distorted results. Multi-group confirmatory factor analysis (MG-CFA) is one of the most commonly used methods to test for MI (Millsap, 2012). However, it is mostly used for comparing two groups. When comparing many groups, the performance of MG-CFA is reduced because the number of measurement parameters to pairwisely compare increases exponentially with the number of groups and non-invariance is falsely detected more easily (Kim, Cao, Wang, & Nguyen, 2017; Rutkowski & Svetina, 2014). Additionally, researchers have to determine the groups or at least the grouping variable a priori (e.g., age or gender; in the following called covariates) (Kim et al., 2017). This often happens with a special application in mind (e.g., cross-cultural comparisons; Milfont & Fischer, 2010) and is mostly done for questionnaires that have already been constructed. We argue that MI should ideally be addressed in the earliest stages of questionnaire development, when changes to the item pool are still easily possible. To address this issue, we want to introduce a method that can help researchers to explore MI in their sample and to automatically identify non-invariant groups: *exploratory factor analysis trees* (*EFA trees*). EFA trees can be seen as an extension of *structural equation model (SEM) trees* introduced by Brandmaier, von Oertzen, McArdle, and Lindenberger (2013b). SEM trees combine SEM with a recursive partitioning algorithm. A SEM is estimated and then tested for parameter instability by a decision tree based method. Thereby, they allow for testing for MI with regard to categorical and continuous covariates (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013a). This is done in a data-driven manner, that is, no covariate has to be chosen in advance. Although decision trees and, thus, SEM trees are already exploratory in nature, so far SEM trees have mainly been applied in the context of CFA or to longitudinal data but to the best of our knowledge not in the context of

EFA (Ammerman, Jacobucci, & McCloskey, 2019; Brandmaier, Driver, & Voelkle, 2018; Brandmaier, Prindle, McArdle, & Lindenberger, 2016; Brandmaier, Ram, Wagner, & Gerstorf, 2017; de Mooij, Henson, Waldorp, & Kievit, 2018; Simpson-Kent et al., 2020; Usami, Hayes, & McArdle, 2017; Usami, Jacobucci, & Hayes, 2019). By introducing EFA trees, we want to extend the SEM tree literature and provide researchers with an easy-to-use method that grasps the full exploratory potential of SEM trees (Goretzko & Bühner, 2022; Jacobucci, Grimm, & McArdle, 2017). We illustrate how EFA trees can be built within the *partykit* R package (Hothorn & Zeileis, 2015) that provides tools for model-based recursive partitioning (Hothorn, Hornik, & Zeileis, 2006; Zeileis, Hothorn, & Hornik, 2008).

The remainder of the paper is structured as follows. First, we describe the concept of MI and its relevance for questionnaire development in more detail. Second, we provide an introduction to EFA. Third, we describe the recursive partitioning algorithm and EFA trees in particular. Last, we show exemplary applications of EFA trees and investigate the performance in identifying a lack of MI under different conditions in simulated examples.

4.3 Measurement Invariance

Assessing MI can be a tedious task. In a factor-analytic framework, four nested levels of MI between groups are considered (Putnick & Bornstein, 2016): a) configural (equal construct architecture; i.e., same number of latent factors and same location of zero loadings in the loading matrices across groups. Note that zero loadings are only imposed in CFA, not in EFA.), b) metric (equal loading sizes), c) scalar (equal intercepts), d) residual (equal unique variance). As already mentioned, MG-CFA is a straightforward way to test for MI (see Putnick & Bornstein, 2016 for an illustrative step-by-step example). However, if there are many groups that have to be compared, this simple approach reaches its limits. The probability of falsely detecting non-invariance increases with number of groups to be compared and model fit might be poor due to strict fit index cut offs (Kim et al., 2017). Rutkowski and Svetina (2014) provided a first remedy to tackle this issue by suggesting adapted cut-offs for model fit measures. Even further, scholars developed other CFA-based methods to test MI in these cases with many groups, for example multilevel factor mixture modelling or alignment optimization

(Asparouhov & Muthén, 2014). Because going into detail about these methods would be beyond the scope of this article, we refer readers interested in CFA-based methods to Kim et al. (2017) for a comprehensive overview. Sass (2011) and Van de Schoot, Lugtig, and Hox (2012) provide general guidelines on testing for MI.

In addition to these CFA-based methods, other EFA-based methods have been developed recently. This resolves some of the aforementioned issues, for example that no restrictive zero loadings have to be imposed. For example, De Roover and Vermunt (2019) developed *multigroup factor rotation* to pinpoint non-invariant loadings between groups. *Mixture multigroup factor analysis* was suggested as a method to cluster groups according to levels of MI, specifically metric (De Roover, Vermunt, & Ceulemans, 2022) and scalar (De Roover, 2021) invariance.

Even though some of these advanced methods can handle many groups, problems arise when there are no particular hypotheses with regard to the covariates defining these groups (Brandmaier et al., 2013b). When there are many covariates (e.g., age, gender, education, ethnicity, etc.), it quickly becomes impossible to test for all of them with all potential group constellations. Usually when researchers test for MI, they define a small number of groups based on one or two covariates (e.g., ethnicity in cross-cultural research). In this, other covariates (or interactions between them) that may define theoretically relevant groups and for which MI cannot be assumed might remain undetected. As Brandmaier et al. (2013b) described, SEM trees can be used to explore the data for non-invariant groups in a data-driven manner (rather than by theoretically deriving hypotheses a priori). Thus, the concept of recursive partitioning seems suitable for exploration of MI with many covariates. To expand this potential to the earliest stages of questionnaire development, we extend SEM trees by EFA trees. Our aim is to add a method to the tool box that can aid researchers in exploring and testing for MI in order to develop questionnaires that considered MI right from the start. Admittedly, this will not render tests for MI prior to actual analyses between two or more defined groups unnecessary. However, EFA trees may improve the measurement quality of psychological constructs and hopefully prevent later issues with data collection and analysis (Jacobucci & Grimm, 2020).

4.4 Exploratory Factor Analysis

EFA is arguably one of the most widely used methods in psychometrics and questionnaire development more specifically. Compared to CFA, there are no constraints on loading paths between the observed variables and the latent factors. Hence, EFA can be used to uncover the relationships between observed and latent variables (Goretzko, Pham, & Bühner, 2021; Mulaik, 2010). More formally, let $\boldsymbol{x} = (x_1, ..., x_p)^\top$ be the p-dimensional vector of observed variables. This vector can be described as a linear function of the m latent factors (Hirose & Yamamoto, 2014; Mulaik, 2010):

$$\boldsymbol{x} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\epsilon} \tag{3}$$

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_p)^\top$ is the p-dimensional vector of intercepts, $\boldsymbol{\Lambda}$ is the $p \times m$ matrix of factor loadings, $\boldsymbol{\xi} = (\xi_1, ..., \xi_m)^\top$ is the m-dimensional vector of latent factor scores, and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_p)^\top$ is the p-dimensional vector of error terms of the observed variables. The error terms are assumed to be normally distributed with mean 0 and variance $\boldsymbol{\Psi}$. $\boldsymbol{\Psi}$ is a $p \times p$ diagonal matrix with the diagonal elements being the unique variances of the observed variables. The factor correlations are captured as the elements of the $m \times m$ matrix $\boldsymbol{\Phi}$. In EFA, the factors have rotational freedom, that is, there exist different sets of factor solutions which have an identical fit to the data but might be easier to interpret. We resolve the issue of rotational freedom by using regularization (an explanation will follow in a later section). The vector $\boldsymbol{x}$ is usually assumed to be multivariate-normally distributed with mean vector $\boldsymbol{\tau}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ (Jöreskog, 1967). In the single-group context, the data are usually standardized so that $\boldsymbol{\tau} = 0$ and $diag(\boldsymbol{\Sigma}) = 1$. In the multi-group context, it is common to keep the data unstandardized and instead use the covariance matrices for model estimation.

We later want to understand how EFA trees detect measurement non-invariance. For this, we have to introduce an estimation function with which the model parameters (i.e., factor loadings, factor correlations, and unique variances) are estimated. The algorithm uses maximum likelihood estimation (MLE). In MLE, parameters are estimated so that the discrepancy between the model-implied covariance matrix $\boldsymbol{\Sigma}$ and the observed

covariance matrix $\boldsymbol{S}$ is minimized (Jöreskog, 1967):

$$F_{MLE}(\boldsymbol{\Sigma}, \mathbf{S}) = \ln|\boldsymbol{\Sigma}| + tr(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}) - \ln|\boldsymbol{S}| - p. \tag{4}$$

MLE has some convenient properties (Fabrigar, Wegener, MacCallum, & Strahan, 1999): In the estimation process, standard errors of the model parameters are computed. These can be used to calculate confidence intervals and assess the statistical significance of factor loadings.[1] Additionally, fit indexes (e.g., RMSEA, CFI, etc.) can be computed that are useful for model evaluation and comparison.

4.5 Score-Based Recursive Partitioning

Now that we have elaborated on how the EFA model is estimated, we turn to the score-based recursive partitioning algorithm (Hothorn et al., 2006; Zeileis et al., 2008). Specifically, how the algorithm finds parameter instability in the model with respect to some covariate and splits the data into heterogeneous groups. The algorithm is based on a tree structure common in machine learning. In detail, the algorithm works as follows (Hothorn et al., 2006; Zeileis et al., 2008):

1. A model (in our case, an EFA) is fit to the entire sample by estimating the model parameters via MLE (see equation (4)). Let $\Pi(\boldsymbol{Y}, \boldsymbol{\theta})$ be the estimation function in equation (4), $\boldsymbol{\theta} = (\boldsymbol{\Lambda}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ the vector of model parameters (i.e., factor loadings, factor correlations, and unique variances) and $\boldsymbol{Y}$ the observations, with elements Y_i, $i = 1, ..., N$. The parameter estimates $\hat{\boldsymbol{\theta}}$ can be obtained by solving the first order condition

$$\sum_{i=1}^{N} \pi(Y_i, \hat{\boldsymbol{\theta}}) = 0 \tag{5}$$

whereby

$$\pi(\boldsymbol{Y}, \boldsymbol{\theta}) = \frac{\partial \Pi(\boldsymbol{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{6}$$

is the score function of $\Pi(\boldsymbol{Y}, \boldsymbol{\theta})$.

[1] To be able to test hypotheses about obliquely rotated factor loadings, Jennrich (1973) showed how to derive the required standard errors.

2. A test for parameter stability is performed with regard to every covariate by means of null hypothesis tests (*structural change test*). For this, we assess whether the corresponding scores evaluated at the parameter estimates, $\hat{\pi}_i = \pi(Y_i, \hat{\boldsymbol{\theta}})$, fluctuate randomly around their mean 0. The scores are ordered according to their deviation from 0 with regard to a covariate. Under the null hypothesis of invariant measurement, the deviations should fluctuate randomly. If, however, the measurement is not invariant, systematic changes in the deviations will be shown by the ordering. The hypothesis tests use different test statistics depending on whether a categorical or continuous covariate is evaluated. In this study, we used a χ^2 test for categorical covariates and the *supLM statistic* (a type of *Lagrange Multiplier statistic*) for continuous covariates. The model needs to be estimated only once to assess MI with regard to different covariates. This is because the amount of score deviations stays the same, only the ordering changes. After every covariate has been evaluated, the one associated with the lowest (Bonferroni-corrected) p-value below a significance level α is selected for splitting the model. Note that by Bonferroni-correcting the p-values, the prespecified significance level α is ensured for the whole tree and the issue of multiple testing is accounted for.
3. Once a covariate for splitting is found, the optimal split point on this covariate has to be computed. When splitting the model into B segments, two potential segmentations can be compared by evaluating the segmented estimation functions $\sum_{b=1}^{B} \sum_{i \in I_b} \Pi(Y_i, \boldsymbol{\theta}_b)$. For continuous covariates, an exhaustive search over all potential segmentations is performed. For a split into $B = 2$ segments, this can be performed in $O(N)$ operations, where N is the sample size. As an example, suppose the continuous variable *age* was identified in step 2 as a covariate that explains parameter instability. To find the optimal split point, the algorithm now loops over every value of *age* from lowest to highest and compares the segmented estimation functions for the groups that would result from splitting at the evaluated value. The value of *age* for which the two segmented estimation functions are optimized is then selected as the split point. For categorical covariates, all potential constellations are evaluated. For a split into $B = 2$ segments, this can be performed in $O(2^{C-1})$ operations, with C

being the number of categories. For example, on the categorical variable *marital status* with four categories, the segmented estimation functions of every group constellation are compared. Again, the constellation for which the estimation functions are optimized is selected for splitting. Theoretically, the model could be split into more than two nodes. However, this diminishes interpretability while simultaneously increasing computational demand (e.g., for continuous variables, a split into more than two groups, $B > 2$, would result in an exhaustive search of order $O(N^{B-1})$). In the following, we only consider the case where the model is split into two nodes (cf. Brandmaier et al., 2013b; Strobl, Kopf, & Zeileis, 2015; Zeileis et al., 2008). Note that if there were three non-invariant groups, they could still be identified by performing binary splits. For this, the algorithm would simply split twice on the same covariate.

4. These steps are repeated until a) no parameter instability in a leaf node becomes statistically significant, b) a prespecified depth of the tree is reached, or c) sample size in a leaf node falls below a prespecified minimal value. For a thorough mathematical introduction see Hothorn et al. (2006), Zeileis and Hornik (2007) and Zeileis et al. (2008).

This algorithm has some convincing advantages (Hothorn et al., 2006; Zeileis et al., 2008): First, it is possible to efficiently test multiple covariates for parameter instability, even without hypotheses about split points. This is especially powerful in the case of continuous covariates like age where manually assessing every potential split point is not feasible (Putnick & Bornstein, 2016). Second, (non-linear) interactions between covariates can be considered. This can be done either by adding the interaction term as a potential covariate or by allowing "deeper" trees. Nodes are conditional on all prior covariates and split points. Hence, in a tree that was split twice on two different covariates, these can be seen as an interaction. Third, the algorithm is unbiased. Other tree algorithms (like CART or C4.5) often tend to favor covariates with many potential split points and are thus biased toward selecting these covariates for splitting. In the score-based recursive partitioning algorithm, this selection bias is eliminated by separating the steps of covariate selection and split point selection. Additionally, the algorithm works on formal parameter stability tests, which also ensures unbiasedness. That is, if the parameters in a node are stable, a false decision to split on any of the

covariates will only be made with a probability of approximately α. Conversely, if the parameters are in fact unstable, and this instability can be explained by a covariate, the instability will be detected for a sufficient sample size N. This is because the tests are consistent at rate $\sqrt{N}$ (Zeileis & Hornik, 2007).

We want to point out that using this recursive partitioning approach is not new in psychometrics and has repeatedly shown good performance. In recent years, it was primarily employed to models in the IRT framework like dichotomous (Strobl et al., 2015) and polytomous (Komboz, Strobl, & Zeileis, 2018) Rasch models. They can be used to detect *differential item functioning* (DIF; Holland & Wainer, 2012) between multiple covariates (Debelak & Strobl, 2019). Schneider, Strobl, Zeileis, and Debelak (2021) provide a tutorial on score-based MI tests in IRT models. We want to extend this literature by combining recursive partitioning with EFA. This might be especially useful for complex constructs where multiple scales ought to be tested for MI simultaneously (Meade & Lautenschlager, 2004). Merkle and Zeileis (2013) and Merkle, Fan, and Zeileis (2014) introduced this algorithm in a factor-analytic context. Their work evaluated the performance of the statistical tests used in our study and thus prepared the technical ground on which our study is built. Both of their studies focused on comparing different test statistics for continuous (Merkle & Zeileis, 2013) and ordinal (Merkle et al., 2014) covariates. We aim to add to this literature by carrying the method to typical psychological research situations. We hope to provide a broader context, for example by considering different violations of MI and types of covariates at the same time. In this, we want to enable substantive researchers to draw on a well-known and commonly used method in psychological questionnaire development when evaluating MI (Fabrigar et al., 1999; Goretzko et al., 2021). This could be especially useful in areas like personality or clinical psychology where constructs are often multi-dimensional.

4.6 EFA Trees

The main purpose of EFA trees is to help researchers to develop questionnaires and psychological tests that have been constructed as measurement invariant as possible. Once a preliminary item set has been built and data have been collected, EFA trees can be used to automatically uncover heterogeneous groups with regard to multiple

covariates. In this, EFA trees can be seen as fully exploratory.

The focus of the succeeding simulations will be on detecting a lack of configural and metric MI (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). That is, we will primarily investigate the ability of EFA trees to assess the construct architecture (i.e., number of latent factors and location of zero loadings) and loading sizes across groups. The two other types of MI, namely scalar (intercept) and residual (unique variances) MI, build on configural and metric MI. Because in EFA data are most often standardized, the mean vector τ becomes 0 and will not be relevant anymore. Additionally, having equal unique variances across groups is hard to achieve and not necessary for a comparison of latent means (Chen, 2007; Putnick & Bornstein, 2016; Vandenberg, 2002).

We investigated whether a lack of MI is actually detected by EFA trees. For this, we first performed four simulations which act as toy examples. In these, we aimed at demonstrating the application and interpretation of EFA trees to questionnaire data. Subsequently, we conducted a comprehensive simulation study in which we manipulated sample size, group size ratio, type of covariate, number of distraction covariates, and type of lack of MI.

4.7 Method

4.7.1 Software

The complete code needed to reproduce all analyses can be found at https://osf.io/7pgrb/. Additionally, we provide a template script which can be used to run an EFA tree with only small adjustments to the code. We conducted all analyses using the statistical software *R* (R Core Team, 2021). The manuscript was written in R markdown using the package *papaja* (Aust & Barth, 2020). We simulated standardized data by drawing from a multivariate normal distribution using the package *mvtnorm* (Genz et al., 2021). The recursive partitioning algorithm was implemented using the package *partykit* (Hothorn & Zeileis, 2015). In the tree growing function *mob*, a control argument can be defined that contains parameters relevant for fitting the algorithm. All control parameters were set as their default values. Most importantly, this means that for the significance level for splitting, we set $\alpha = 0.05$ and p-values were Bonferroni-corrected. For all analyses, we specified a three-dimensional model with 18 observed variables using the package *lavaan* (Rosseel, 2012). Every observed variable was allowed to load freely on every factor. Because the recursive partitioning algorithm cannot handle unidentified models, we first defined a model with uncorrelated factors to ensure identification. By setting the argument *auto.efa = TRUE* in the *lavaan* function, all constraints to identify a model were imposed: factor correlations were set to 0, factor variances were set to 1, and some factor loadings were constrained to followed an echelon pattern (Rosseel, 2012). Because we assume that no information about the items or the data is available in advance, it is difficult to provide a general recommendation regarding the selection of loadings to constrain. If the wrong loadings are constrained, parameter differences that are critical for the assessment of non-invariance might remain undetected. However, one can empirically assess whether different selections of constrained loadings have a considerable influence (Dolan, Oort, Stoel, & Wicherts, 2009). This can be done by growing more than one tree in parallel with different constrained loadings and comparing the results.

4.7.2 Toy examples

4.7.2.1 Procedure.

The algorithm was employed as described in the section *Software*. To demonstrate exemplary applications and interpretations of EFA trees, we further investigated the estimated models in the leaf nodes after splitting. For this, we extracted the data from these nodes and re-estimated an unidentified model with correlated factors and all loading paths freed using regularized EFA (Hirose & Yamamoto, 2014; Scharf & Nestler, 2019). We want to briefly explain our rationale behind using regularized EFA: Once the EFA model has been estimated, researchers often aim at obtaining an interpretable solution of the matrix of factor loadings $\mathbf{\Lambda}$ (Mulaik, 2010). The most common goal is to achieve a so-called simple structure. That is, each item has one high loading on one factor and low to no cross-loadings on all other factors. The method of choice to obtain such a structure is rotation of factor solutions. EFA models are rotationally indeterminate, that is, there is an infinite set of factor solutions that fits a data set equally well (Mulaik, 2010). Many rotation methods exist with no one best method (Browne, 2001; Trendafilov, 2014). The best method to use in a specific application depends on the true factor structure in the population. Because this population factor structure is almost always unknown, the choice of rotation method is rather subjective (Asparouhov & Muthén, 2009; Sass & Schmitt, 2010; trying different hyperparameter settings of the *simplimax* rotation could help to find a solution with most loadings close to zero; see Kiers, 1994 for more details).

The very goal of EFA trees is to uncover different structures of a construct between groups. Thus, it is difficult to pick an optimal rotation method for every EFA estimated in a leaf node of the resulting tree.[2] Taking this into account, we applied regularized EFA to obtain interpretable factor solutions in the leaf nodes (Hirose & Yamamoto, 2014; Jacobucci, Grimm, & McArdle, 2016). As Scharf and Nestler (2019) demonstrated in a comprehensive comparison of common rotation methods and regularization, the latter is not necessarily "better" than rotation in recovering simple structure. However,

[2]An interesting extension could be to combine EFA trees with the aforementioned multigroup factor rotation (MGFR; De Roover & Vermunt, 2019). Instead of regularizing the models in the nodes, MGFR could be applied to investigate group-specific measurement models in the leaf nodes. One advantage of this approach over regularization would be that one could pinpoint the parameters that differ across the nodes.

it proves more objective in the sense that the true structure of the construct does not have to be known. Essentially, regularization switches the rotation problem of EFA to a variable selection problem. The regularization was implemented using the package *regsem* (Jacobucci et al., 2016). We used elastic net regularization (Zou & Hastie, 2005) and penalized both the factor loadings and the factor correlations. The hyperparameters γ (controlling the amount of regularization) and β (controlling the type of regularization) were tuned by choosing values that minimized the BIC over the whole sample (Jacobucci et al., 2016). For γ, we tested 100 values in a grid search starting from $\gamma = 0.001$ with a step size of 10^{-5}. For β, we tested all values between 0.05 and 0.95 with a step size of 0.05 (cf. Scharf & Nestler, 2019). For *regsem*, an unidentified model was not an issue because the *cv_regsem* function only requires the model-implied covariance matrix, not an identified model. In the process of estimation, the model eventually became identified due to variable selection (Li, Jacobucci, & Ammerman, 2021).

4.7.2.2 Toy Example 1: Configural Invariance - Different Number of Factors.

In a first toy example, we investigated whether an EFA tree would detect a violation of configural invariance caused by differing numbers of latent factors between groups. Suppose our construct that was measured by 18 indicators. For men, these indicators were described by three latent factors, whereas for women, there were four latent factors. The standardized loading matrices on population level were (cf. Scharf & Nestler, 2019):

$$\mathbf{\Lambda_{Men}} = \begin{bmatrix} 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \end{bmatrix}; \mathbf{\Lambda_{Women}} = \begin{bmatrix} 0.75 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 \\ 0.75 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.75 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0 & 0 & 0.75 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0.75 & 0 \\ 0 & 0 & 0 & 0.75 \end{bmatrix}$$

As can be seen, the loading matrices of both men and women did not have cross-loadings. However, the last indicator of each of the three factors in the group of men was shifted to a forth factor in the group of women. We simulated a data set with $N = 400$ and the dichotomous covariate *sex*, consisting of 200 men and 200 women. Additionally, we simulated four covariates as "distractors" to mimic a setting typical for questionnaire development: two (standard-normally distributed) continuous, one other dichotomous, and one categorical covariate with four categories. These covariates were independent from the factorial structure on population level but could have potentially been selected by the EFA tree as a split variable. As described above, we estimated a model with three factors (i.e., a misspecified model for women). In parametric notation, the present violation of configural invariance means that Φ is a 3×3 matrix for men and a 4×4 matrix for women.

The results of the analysis are shown in Table 1. The EFA tree successfully identified the covariate *sex* for splitting and ignored the four other covariates. Thus, all men and all women ended up in two different leaf nodes. We also conducted a parallel analysis (Horn, 1965) on the data in each leaf node, which correctly suggested three factors in the "male node" and four factors in the "female node." Especially in the early stages of questionnaire development, a parallel analysis in each leaf node seems beneficial.

Table 1
Study 1: Test statistics and p-values for toy example 1

	sex	cov1	cov2	cov3	cov4
statistic	152.45	80.20	122.47	57.43	208.36
p.value	0.00	1.00	0.02	1.00	0.95

Note. Test statistics were a χ^2 test for categorical and the supLM statistic for continuous covariates. cov1 - cov4 denote the distractor covariates.

Table 2 shows the loading matrices of the regularized EFA models in the two leaf nodes. The matrices with the theoretically assumed three latent factors show no clear cause of the violation of MI. However, in the matrix with four latent factors in the female node, as indicated by the data, it can be seen that the observed variables 6, 12, and 18 load on an additional factor not present in the male node. Unfortunately, different number of factors cannot be evaluated directly because the algorithm can only handle one pre-specified model. However, in these cases it would remain unclear anyways what different numbers of latent factors mean on a conceptual level. This emphasizes that further analyses on the data in the leaf nodes are crucial to better understand your data and, ultimately, the construct of interest (cf. Brandmaier et al., 2013b).

Table 2
Study 1: Regularized factor solution for toy example 1

Men			Women 3 Factors			Women 4 Factors			
F1	F2	F3	F1	F2	F3	F1	F2	F3	F4
0.72	0.01	0.13	0.73	0.08	0.03	0.72	0.09	0.00	0.03
0.72	0.04	0.07	0.64	0.09	0.03	0.63	0.10	0.01	0.02
0.75	0.00	0.01	0.76	0.00	0.00	0.76	0.00	0.06	-0.02
0.80	0.07	0.01	0.78	-0.03	0.06	0.77	-0.01	-0.02	0.06
0.73	0.06	-0.01	0.68	0.07	0.10	0.67	0.07	0.03	0.10
0.76	0.00	0.00	0.04	0.19	0.23	0.03	0.04	0.68	0.09
0.11	0.72	0.07	0.14	0.72	0.03	0.14	0.72	0.02	0.04
0.00	0.73	0.11	-0.03	0.70	0.10	-0.03	0.71	0.00	0.11
0.11	0.71	-0.04	0.05	0.77	0.04	0.05	0.78	-0.01	0.05
0.04	0.75	0.07	0.06	0.70	0.00	0.06	0.69	0.04	0.00
0.10	0.75	0.00	0.00	0.73	-0.02	0.00	0.69	0.15	-0.04
-0.05	0.77	0.15	0.00	0.22	0.16	0.00	0.07	0.72	0.00
0.13	0.00	0.69	0.00	0.02	0.73	0.00	0.02	0.00	0.74
0.02	0.06	0.73	-0.03	0.00	0.83	-0.03	-0.01	0.08	0.81
0.00	0.16	0.72	0.02	0.05	0.74	0.02	0.06	-0.02	0.75
0.00	0.09	0.71	0.00	0.13	0.75	0.00	0.12	0.07	0.73
0.04	-0.01	0.76	0.02	0.00	0.73	0.02	0.00	0.04	0.72
0.02	0.13	0.71	0.01	0.16	0.26	0.00	0.00	0.70	0.12

Note. F1 - F4 denote the latent factors. The factor solutions were achieved by re-estimating the models in the leaf nodes via elastic net regularization.

4.7.2.3 Toy Example 2: Configural Invariance - Simple Structure vs. Cross-Loadings.

In the second toy example, we looked at a different form of configural non-invariance, that is, simple structure in one group and cross-loadings in the other group. We again used the construct with 18 indicators from toy example 1. This time, the number of latent factors was three for both groups. However, the two groups were now defined by a (standardized) continuous covariate *age*. We simulated the groups based on the z-scores at the mean 0: $z_{age} \leq 0$ was the "younger" group and $z_{age} > 0$ the "older" group. This yielded approximately equally sized groups. Note that while this leads to two age-groups that have to be uncovered by the EFA tree, it still has to treat *age* as a continuous variable when assessing parameter instability on this covariate. The standardized loading matrix on population level for the younger group was the same as the one of the men used in toy example 1. For the older group, cross-loadings were added (cf. Scharf & Nestler, 2019):

$$\mathbf{\Lambda_{Younger}} = \begin{bmatrix} 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \end{bmatrix}; \mathbf{\Lambda_{Older}} = \begin{bmatrix} 0.67 & 0.22 & 0.13 \\ 0.68 & 0.09 & 0.23 \\ 0.68 & 0.27 & 0.05 \\ 0.65 & 0.39 & 0.09 \\ 0.64 & 0.13 & 0.39 \\ 0.67 & 0.18 & 0.18 \\ 0.05 & 0.68 & 0.27 \\ 0.28 & 0.63 & 0.38 \\ 0.38 & 0.63 & 0.21 \\ 0.09 & 0.69 & 0.18 \\ 0.05 & 0.73 & 0.05 \\ 0.27 & 0.67 & 0.13 \\ 0.04 & 0.40 & 0.66 \\ 0.38 & 0.25 & 0.63 \\ 0.26 & 0.18 & 0.66 \\ 0.14 & 0.09 & 0.70 \\ 0.22 & 0.22 & 0.66 \\ 0.18 & 0.09 & 0.69 \end{bmatrix}$$

Λ_{Older} had its main loadings at the same location as $\Lambda_{Younger}$ but had (considerable) cross-loadings (up to 0.4). We simulated a data set with $N = 1000$ and the continuous covariate *age* that defined the two groups as described above. Again, we simulated four distractors: two other (standard-normally distributed) continuous, one dichotomous, and one categorical covariate with four categories. Factor correlations on population level were 0.3, factor variances were fixed to 1.

The results are shown in Table 3. The EFA tree identified the covariate *age* and split the data approximately at $z_{age} = 0$ (one observation from the younger group near the mean 0 was falsely put in the leaf node of the older group). It ignored all other covariates. The standardized root mean square residuals (SRMR) of the EFA models in the leaf nodes were 0.01 and 0.01 for the younger and the older group, respectively, indicating good fit. Additionally, regularization of the models in the leaf nodes (approximately) recovered the simple structure in the younger group and the considerable cross-loadings of some observed variables in the older group (see Table 4). This could be considered an indication of configural non-invariance.

Table 3
Study 1: Test statistics and p-values for toy example 2

	age	cov1	cov2	cov3	cov4
statistic	474.38	94.96	92.62	67.93	217.60
p.value	0.00	0.90	0.95	0.97	0.82

Note. Test statistics were a χ^2 test for categorical and the supLM statistic for continuous covariates. cov1 - cov4 denote the distractor covariates.

It should be mentioned here that even though an EFA tree can efficiently test parameter stability on a continuous covariate, in the end it still makes a binary split. While this might fail to capture gradual differences in parameters, it has the advantage of interpretability. If one is willing to make the assumption that there are two discrete groups that are defined along a continuous covariate, EFA trees yield two fully interpretable and employable models. Additionally, one does not have to prespecify any covariates that might be associated with non-invariance in the data. We refer readers who want to assess gradual parameter differences along a known continuous covariate (without having to split the data) to literature on *multiple indicator multiple cause models* (*MIMIC models*; Muthén, 1989). Note, however, that this approach does not use fewer assumptions. For example, one assumption that is as strict as ours of two discrete groups is the exact functional form of gradual differences included in a MIMIC model (i.e., linear/quadratic/...).

Table 4
Study 1: Regularized factor solution for toy example 2

Younger			Older		
F1	F2	F3	F1	F2	F3
0.76	-0.01	0.01	0.70	0.11	0.11
0.76	0.00	0.00	0.65	0.00	0.29
0.72	0.04	0.06	0.72	0.20	0.00
0.76	0.06	-0.02	0.69	0.30	0.06
0.76	0.04	0.02	0.62	0.00	0.42
0.75	0.11	0.04	0.68	0.11	0.15
-0.06	0.78	0.04	0.00	0.66	0.26
0.01	0.77	0.01	0.27	0.59	0.36
0.00	0.78	0.06	0.42	0.57	0.17
0.01	0.73	0.05	0.11	0.64	0.16
0.05	0.76	0.00	0.05	0.78	-0.06
0.03	0.76	-0.03	0.28	0.65	0.07
0.01	0.00	0.72	-0.03	0.35	0.68
0.04	0.00	0.74	0.31	0.15	0.68
0.05	0.00	0.76	0.19	0.13	0.68
-0.01	0.03	0.72	0.08	0.00	0.76
0.00	0.06	0.77	0.15	0.09	0.74
0.02	0.00	0.78	0.08	0.04	0.75

Note. F1 - F3 denote the latent factors. The factor solutions were achieved by re-estimating the models in the leaf nodes via elastic net regularization.

4.7.2.4 Toy Example 3: Metric Invariance - Different Loading Sizes.

In a third toy example, metric invariance of our three-dimensional construct with 18 indicators was violated by a categorical covariate *marital status* with four categories. More specifically, loading sizes are different for observations that are "single" from observations from all other categories. The standardized loading matrix on population level for single observations was the same as the one of the older group used in toy example 2. For all other categories, cross-loadings were noticibly smaller (cf. Scharf & Nestler, 2019):

$$\mathbf{\Lambda_{Single}} = \begin{bmatrix} 0.67 & 0.22 & 0.13 \\ 0.68 & 0.09 & 0.23 \\ 0.68 & 0.27 & 0.05 \\ 0.65 & 0.39 & 0.09 \\ 0.64 & 0.13 & 0.39 \\ 0.67 & 0.18 & 0.18 \\ 0.05 & 0.68 & 0.27 \\ 0.28 & 0.63 & 0.38 \\ 0.38 & 0.63 & 0.21 \\ 0.09 & 0.69 & 0.18 \\ 0.05 & 0.73 & 0.05 \\ 0.27 & 0.67 & 0.13 \\ 0.04 & 0.40 & 0.66 \\ 0.38 & 0.25 & 0.63 \\ 0.26 & 0.18 & 0.66 \\ 0.14 & 0.09 & 0.70 \\ 0.22 & 0.22 & 0.66 \\ 0.18 & 0.09 & 0.69 \end{bmatrix}; \mathbf{\Lambda_{Rest}} = \begin{bmatrix} 0.70 & 0.11 & 0.14 \\ 0.70 & 0.17 & 0.05 \\ 0.68 & 0.16 & 0.16 \\ 0.70 & 0.05 & 0.17 \\ 0.72 & 0.08 & 0.08 \\ 0.70 & 0.11 & 0.11 \\ 0.11 & 0.69 & 0.17 \\ 0.05 & 0.72 & 0.08 \\ 0.05 & 0.72 & 0.08 \\ 0.16 & 0.68 & 0.16 \\ 0.08 & 0.71 & 0.11 \\ 0.05 & 0.71 & 0.14 \\ 0.08 & 0.14 & 0.70 \\ 0.14 & 0.14 & 0.69 \\ 0.14 & 0.11 & 0.70 \\ 0.11 & 0.05 & 0.71 \\ 0.16 & 0.14 & 0.69 \\ 0.08 & 0.05 & 0.72 \end{bmatrix}$$

Cross-loadings of Λ_{Single} were as high as 0.40, whereas in Λ_{Rest} they reached a maximum of 0.17. We simulated a data set with $N = 400$ and the categorical covariate *marital status*. In *marital status*, each category had $n = 100$ observations. This time, we simulated eight distractors: four (standard-normally distributed) continuous, two dichotomous, one other categorical with four categories, and one ordinal covariate with four categories. Factor correlations on population level were 0.3, factor variances were fixed to 1.

The results are shown in Table 5. The EFA tree split the data into single and non-single observations. Every observations was put in the correct leaf node and no distractor was chosen for splitting. The SRMRs of the EFA models in the leaf nodes were 0.03 and 0.02 for the singles and the rest group, respectively, indicating good fit.

Table 5
Study 1: Test statistics and p-values for toy example 3

	marital status	cov1	cov2	cov3	cov4	cov5	cov6	cov7	cov8
statistic	276.70	92.67	87.40	85.69	82.11	70.32	66.77	222.31	215.59
p.value	0.01	0.99	1.00	1.00	1.00	0.99	1.00	0.89	0.97

Note. Test statistics were a χ^2 test for categorical and the supLM statistic for continuous covariates. cov1 - cov8 denote the distractor covariates.

Further inspection of the models in the leaf nodes showed that the recovery of the population loading matrices was not perfect (see Table 6). It is important to consider

that regularization might yield imperfect solutions, for example if some parameters are shrunk too much toward zero. However, in our fully exploratory setting, one can still see that cross-loadings differ in their amount between the two groups, suggesting metric non-invariance.

Table 6

Study 1: Regularized factor solution for toy example 3

Single			Rest		
F1	F2	F3	F1	F2	F3
0.70	0.02	0.10	0.73	0.04	0.10
0.64	0.00	0.30	0.77	0.04	0.00
0.76	0.12	-0.03	0.72	0.00	0.15
0.70	0.35	0.01	0.75	0.04	0.12
0.62	-0.01	0.45	0.69	0.00	0.09
0.76	0.11	0.00	0.73	0.00	0.12
-0.05	0.80	0.11	0.08	0.75	0.14
0.23	0.62	0.32	0.02	0.69	0.09
0.37	0.67	0.05	0.00	0.77	0.00
0.04	0.80	0.05	0.07	0.72	0.13
0.00	0.75	-0.03	0.10	0.70	0.05
0.26	0.69	0.00	0.06	0.71	0.06
-0.10	0.37	0.70	0.00	0.19	0.70
0.31	0.11	0.72	0.07	0.09	0.71
0.24	0.14	0.63	0.06	0.12	0.73
0.00	0.06	0.78	0.07	0.06	0.69
0.05	0.19	0.72	0.12	0.13	0.68
0.08	0.00	0.75	0.04	-0.06	0.79

Note. F1 - F3 denote the latent factors. The factor solutions were achieved by re-estimating the models in the leaf nodes via elastic net regularization.

4.7.2.5 Toy Example 4: Configural and Metric Invariance - Interaction effects between covariates.

In a forth toy example, we investigated whether EFA trees can capture interaction effects between covariates. Recall that interactions can be detected by allowing the tree to split more than once. If a tree subsequently splits data on two different covariates, these splits can be seen as an interaction between the two split covariates. Again, we assume our three-dimensional construct with 18 indicators. MI was violated by a categorical variable *sex* in that the population loading matrix for women showed a perfect simple structure whereas for men, cross-loadings were present (i.e., a violation of configural MI). Additionally, in the "male" leaf node, the population matrices for men above and below the mean age differed with respect to the size of the cross-loadings

(i.e., a violation of metric MI; cf. toy example 2). That is, there was an interaction effect between *sex* and *age* in the sense that only the population matrices of men were affected by *age*. The standardized loading matrices on population level were the same as in toy example 2 and 3 (cf. Scharf & Nestler, 2019):

$$\mathbf{\Lambda_{Women}} = \begin{bmatrix} 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0.75 & 0 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0.75 & 0 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \\ 0 & 0 & 0.75 \end{bmatrix}$$

$$\mathbf{\Lambda_{Younger\ men}} = \begin{bmatrix} 0.67 & 0.22 & 0.13 \\ 0.68 & 0.09 & 0.23 \\ 0.68 & 0.27 & 0.05 \\ 0.65 & 0.39 & 0.09 \\ 0.64 & 0.13 & 0.39 \\ 0.67 & 0.18 & 0.18 \\ 0.05 & 0.68 & 0.27 \\ 0.28 & 0.63 & 0.38 \\ 0.38 & 0.63 & 0.21 \\ 0.09 & 0.69 & 0.18 \\ 0.05 & 0.73 & 0.05 \\ 0.27 & 0.67 & 0.13 \\ 0.04 & 0.40 & 0.66 \\ 0.38 & 0.25 & 0.63 \\ 0.26 & 0.18 & 0.66 \\ 0.14 & 0.09 & 0.70 \\ 0.22 & 0.22 & 0.66 \\ 0.18 & 0.09 & 0.69 \end{bmatrix}; \mathbf{\Lambda_{Older\ men}} = \begin{bmatrix} 0.70 & 0.11 & 0.14 \\ 0.70 & 0.17 & 0.05 \\ 0.68 & 0.16 & 0.16 \\ 0.70 & 0.05 & 0.17 \\ 0.72 & 0.08 & 0.08 \\ 0.70 & 0.11 & 0.11 \\ 0.11 & 0.69 & 0.17 \\ 0.05 & 0.72 & 0.08 \\ 0.05 & 0.72 & 0.08 \\ 0.16 & 0.68 & 0.16 \\ 0.08 & 0.71 & 0.11 \\ 0.05 & 0.71 & 0.14 \\ 0.08 & 0.14 & 0.70 \\ 0.14 & 0.14 & 0.69 \\ 0.14 & 0.11 & 0.70 \\ 0.11 & 0.05 & 0.71 \\ 0.16 & 0.14 & 0.69 \\ 0.08 & 0.05 & 0.72 \end{bmatrix}$$

We simulated a data set with $N = 1000$ together with the categorical covariate *sex* and the continuous covariate *age*. In *sex*, there were $n = 300$ women and $n = 700$ men. Of these 700 men, $n = 354$ were younger than the mean age and $n = 346$ were older. We simulated four distractors: two (standard-normally distributed) continuous covariates, one dichotomous covariate, and one categorical covariate with four categories. Factor correlations on population level were 0.3, factor variances were fixed to 1.

Table 7
Study 1: Test statistics and p-values for the first node in toy example 4

	sex	age	cov1	cov2	cov3	cov4
statistic	269.00	235.69	81.69	106.02	85.48	187.68
p.value	0.00	0.00	1.00	0.43	0.42	1.00

Note. Test statistics were a χ^2 test for categorical and the supLM statistic for continuous covariates. cov1 - cov4 denote the distractor covariates.

Table 8
Study 1: Test statistics and p-values for the second node in toy example 4

	sex	age	cov1	cov2	cov3	cov4
statistic	0.00	328.80	87.54	100.67	88.89	195.11
p.value	NA	0.00	0.99	0.68	0.24	1.00

Note. Test statistics were a χ^2 test for categorical and the supLM statistic for continuous covariates. cov1 - cov8 denote the distractor covariates.

The results are shown in Table 7 (for the first node) and Table 8 (for the second node). Note that the p-values for both covariates *sex* and *age* were below the Bonferroni-correct level of significance of 0.05 but the p-value of *sex* was lower than that of *age*. Thus, the EFA tree first split the data on the covariate *sex*. Subsequently, it performed a second split on the covariate *age* in the male leaf node at $z_{age} = -0.00455$. This split point was not exactly optimal because it led two observations that had values below the mean 0 but above the split point ($-0.00455 < z_{age} < 0$) to falsely end up in the "older male" leaf node. Nonetheless, the EFA tree correctly identified the interaction effect between *sex* and *age*. The SRMRs of the EFA models in the leaf nodes were 0.02, 0.01, and 0.02 for the female, the younger male, and the older male groups, respectively. Further inspection of the models in the leaf nodes showed an approximate simple structure for women and cross-loadings for men, with high cross-loadings for younger men and rather small cross-loadings for older men (see Table 9).

Table 9
Study 1: Regularized factor solution for toy example 4

Women			Younger men			Older men		
F1	F2	F3	F1	F2	F3	F1	F2	F3
0.74	0.04	0.02	0.74	0.16	0.01	0.76	0.09	0.00
0.70	0.00	0.14	0.75	-0.04	0.15	0.71	0.20	0.00
0.71	0.05	0.00	0.75	0.16	0.00	0.72	0.11	0.07
0.77	0.09	0.00	0.72	0.31	-0.08	0.77	0.00	0.08
0.72	0.01	0.13	0.68	0.05	0.30	0.76	0.02	0.06
0.76	-0.02	0.02	0.80	0.00	0.13	0.66	0.16	0.02
0.02	0.76	0.00	0.00	0.61	0.34	0.10	0.68	0.15
0.00	0.72	0.04	0.28	0.54	0.36	-0.03	0.72	0.06
0.10	0.66	0.07	0.42	0.56	0.13	0.02	0.75	0.00
0.01	0.74	0.12	0.18	0.61	0.15	0.15	0.69	0.08
-0.02	0.71	0.08	0.09	0.72	0.00	0.00	0.76	0.00
0.04	0.71	0.00	0.26	0.64	0.08	0.00	0.80	0.04
0.04	0.13	0.69	-0.06	0.35	0.72	0.16	0.12	0.66
0.00	0.10	0.71	0.42	0.12	0.59	0.08	0.16	0.69
0.03	0.08	0.67	0.34	0.06	0.62	0.16	0.03	0.71
0.03	0.00	0.72	0.09	0.00	0.77	0.12	0.00	0.70
0.04	-0.04	0.76	0.24	0.10	0.67	0.17	0.12	0.64
-0.03	0.07	0.72	0.21	0.04	0.64	-0.06	0.00	0.81

Note. F1 - F3 denote the latent factors. The factor solutions were achieved by re-estimating the models in the leaf nodes via elastic net regularization.

In summary, the toy examples showed that EFA trees can uncover a lack of MI under typical questionnaire research conditions. One of the main advantages of the method is that it allows substantive researchers to do what they are used to. They estimate an EFA and interpret factor loadings by investigating the content of different items and by making sense of latent factors. The only difference is that now researchers get to work with two (or possibly more) loading matrices, being able to better understand heterogeneous groups in their data. However, you do not get statistical information on which parameters differ across the nodes. This highlights the need for thorough investigations of the models in the leaf nodes with domain expertise.[3] As already mentioned, an interesting future extension would be to combine EFA trees with MGFR (De Roover & Vermunt, 2019) to identify specific parameters differences. In the following, we report the results of a structured simulation study to investigate the performance of the trees under various conditions.

[3]During the review process, one reviewer posed the question whether EFA trees would also split the data if differences occured only in factor correlations between groups. We have created an online supplement in which we show that EFA trees split the data in this case and demonstrate what this entails for the invariance of measurements. Additionally, we discuss the use of covariance instead of correlation matrices when estimating the models in the leaf nodes. The online supplement is openly available at https://osf.io/7pgrb/.

4.8 Simulation Study

4.8.1 Procedure

The algorithm was employed as described in the section *Software*. The simulation study was run on the Linux-cluster of the Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities. We manipulated five variables that mimic typical research conditions and could potentially influence the performance of the trees:

- *Sample size*: 400 vs. 1,000 vs. 10,000. With sample sizes of 400 and 1,000 we investigated conditions typical for questionnaire research (Fabrigar et al., 1999; Goretzko et al., 2021) and with a sample size of 10,000 we investigated the asymptotic properties of EFA trees.
- *Type of split covariate*: categorical vs. continuous. The split variable was either a categorical (binary) or a continuous variable (following a standard-normal distribution).
- *Group size ratio*: 50/50 vs. 20/80. The group sizes in the leaf nodes were either equal, or skewed so that 20% of the whole sample belonged to one leaf node and 80% belonged to the other one. For some conditions with a continuous split covariate, these ratios were only approximately achieved due to random number generation from a normal distribution. That is, data for the covariate were first drawn randomly from a standard-normal distribution and were then split into two groups by choosing a cut point that would lead to the desired group size ratios (cf. toy example 2). For example, for the ratio 50/50 that corresponded to a cut point at $z = 0$. Whereas in theory, this should divide the sample into two equally sized groups, in practice it could happen that the ratio is not exactly 50/50 because out of the N observations, a few more might have been generated on one side of the cut point than on the other.
- *Number of distractor covariates*: 4 vs. 8. For the condition with four distractors, we simulated one (standard-normally distributed) continuous covariate, one binary covariate, one categorical covariate with four categories, and one ordinal covariate with four categories. For the condition with eight distractors, we simulated two of these covariates each.

- *Type of lack of MI*: configural vs. metric. In the condition with lack of configural MI, we used the loading matrices from toy example 2 (simple structure vs. cross-loadings). In the condition with lack of metric MI, we used the loading matrices from toy example 3 (small cross-loadings vs. considerable cross-loadings).

We refrained from including conditions in which the covariates are correlated. This is a rather simplified setting, but our goal was to provide a first large-scale simulation to show the performance of model-based recursive partitioning in combination with EFA. In future studies, we plan to investigate the performance of EFA trees under more nuanced conditions; e.g., U-shaped relations between parameter instability and covariates, complicated interactions, and also correlated covariates.

We also added six conditions in which MI was supported, i.e. in which EFA trees should not split the data (3 sample sizes $\times$ 2 numbers of distractors). In total, this amounted to 54 conditions. We simulated 1,000 data sets per condition, resulting in 54,000 data sets for the analysis. As dependent variables, we compared the type I error rates (i.e., the rate of falsely splitting invariant data) and type II error rates (i.e., the rate of falsely missing a split of non-invariant data). Additionally, we looked at the mean and standard deviation (SD) of the SRMR in the leaf nodes.

4.8.2 Results

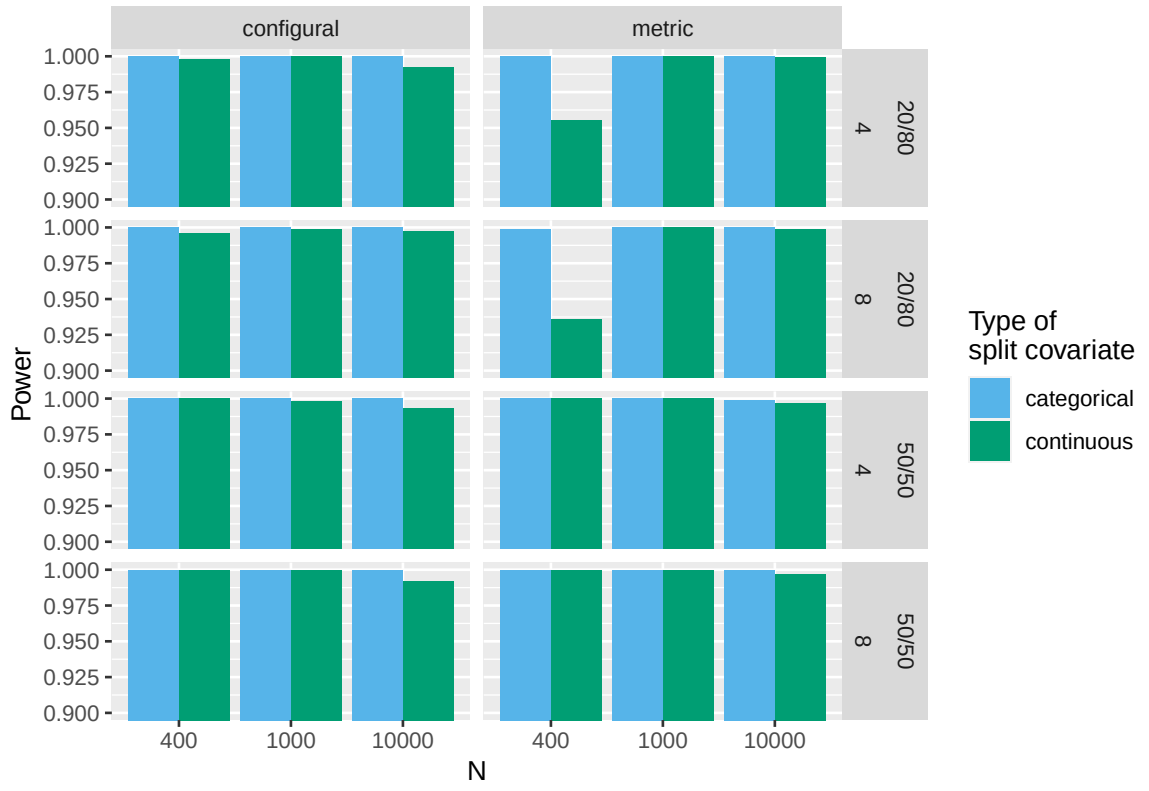

Figure 1
Study 1: Power (1 - type II error rate) of EFA trees to detect lack of measurement invariance (MI) by sample size N. Configural and metric denote the type of lack of MI. 20/80 and 50/50 denote the group size ratio. 4 and 8 denote the number of distractors.

Figure 1 shows the power (i.e., the rate of correctly detecting a lack of MI; 1 - type II error rate) of EFA trees for all conditions. Overall, EFA trees demonstrated a high power of $> 93\%$ for all conditions. EFA trees only missed a split in conditions where sample size was 400; for the conditions of sample size 1,000 and 10,000 the data was always split. However, in rare occasions for sample sizes 1,000 and 10,000, EFA trees chose the wrong covariate for splitting and then encountered problems of estimating the EFA models in the leaf nodes. We assume that this was due to too few observations in the nodes after a wrong split covariate (and thus, a wrong split point) was chosen. For the two conditions of sample size 400, ratio between groups 20/80, continuous split covariate, lack of metric MI, and number of distractors four and eight (ceteris paribus) the power was markedly smaller than for all other conditions (95.5% and 93.6%, respectively). Nonetheless, the power for these conditions can still

be considered good and they are arguably the most complex conditions (small sample size, unbalanced groups, continuous covariate, and comparison of different sizes of cross-loadings).

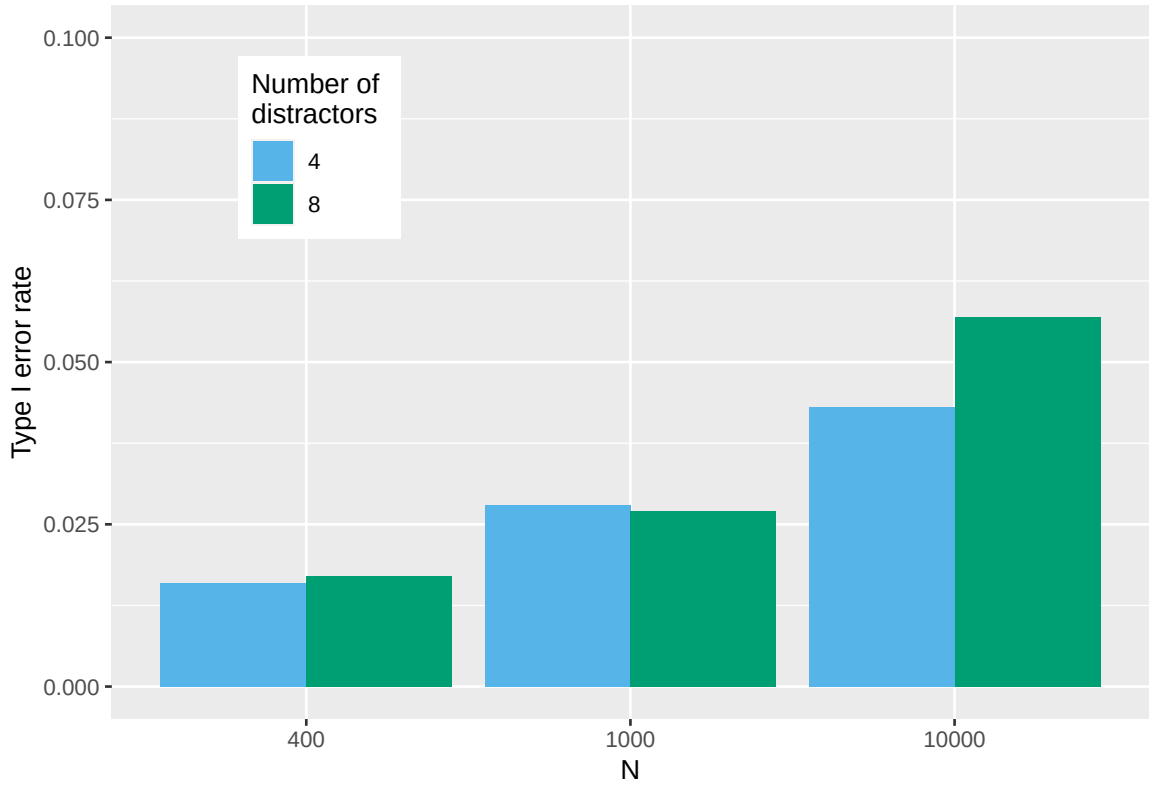

Figure 2
Study 1: Type I error rate (false-positive rate) of EFA trees by sample size N and number of distractors.

Figure 2 shows the type I error rates by sample size and by number of covariates. Most notably, the rate increased with sample size. The type I error rates did not markedly exceed the significance level we set for the EFA trees ($\alpha = 0.05$). Only for a sample size of 10,000 and eight distractors the observed type I error rate was higher (0.057). When constructing an approximate 95% Wald-confidence interval (CI) around the observed type I error rates, the CI for sample size 10,000 contained the nominal level of significance $\alpha = 0.05$. However, for sample sizes 400 and 1,000 it did not contain 0.05. This could be an indication that the parameter stability tests are overly conservative. While from a statistical point of view this might not be ideal, the power to detect non-invariance was still high in our study. Nonetheless, future simulations should investigate the behavior of the type I error rate with even larger sample sizes or

different test statistics.

Table 10

Study 1: Mean and standard deviations of the standardized root mean squared residuals in the two leaf nodes and split rates for all 54 conditions

Condition	Child Node 1	Child Node 2	SD1	SD2	Split Rate
1000, 4, 20/80, categorical, configural	0.016	0.012	0.0014	0.0014	1.000
1000, 4, 20/80, categorical, metric	0.015	0.010	0.0015	0.0012	1.000
1000, 4, 20/80, continuous, configural	0.012	0.016	0.0009	0.0018	1.000
1000, 4, 20/80, continuous, metric	0.010	0.016	0.0008	0.0018	1.000
1000, 4, 50/50, categorical, configural	0.010	0.015	0.0008	0.0013	1.000
1000, 4, 50/50, categorical, metric	0.010	0.013	0.0008	0.0012	1.000
1000, 4, 50/50, continuous, configural	0.015	0.010	0.0012	0.0024	0.998
1000, 4, 50/50, continuous, metric	0.013	0.010	0.0011	0.0026	1.000
1000, 4, none, none, none	0.007	0.010	0.0013	0.0031	0.028
1000, 8, 20/80, categorical, configural	0.016	0.012	0.0014	0.0013	1.000
1000, 8, 20/80, categorical, metric	0.015	0.010	0.0014	0.0008	1.000
1000, 8, 20/80, continuous, configural	0.012	0.016	0.0009	0.0017	0.999
1000, 8, 20/80, continuous, metric	0.010	0.015	0.0008	0.0016	1.000
1000, 8, 50/50, categorical, configural	0.010	0.015	0.0008	0.0012	1.000
1000, 8, 50/50, categorical, metric	0.010	0.013	0.0008	0.0014	1.000
1000, 8, 50/50, continuous, configural	0.015	0.010	0.0012	0.0021	1.000
1000, 8, 50/50, continuous, metric	0.013	0.010	0.0010	0.0022	1.000
1000, 8, none, none, none	0.007	0.012	0.0010	0.0053	0.027
10000, 4, 20/80, categorical, configural	0.005	0.004	0.0004	0.0009	1.000
10000, 4, 20/80, categorical, metric	0.005	0.003	0.0004	0.0007	1.000
10000, 4, 20/80, continuous, configural	0.004	0.005	0.0003	0.0018	0.992
10000, 4, 20/80, continuous, metric	0.003	0.005	0.0002	0.0005	0.999
10000, 4, 50/50, categorical, configural	0.003	0.005	0.0002	0.0005	1.000
10000, 4, 50/50, categorical, metric	0.003	0.004	0.0002	0.0006	0.999
10000, 4, 50/50, continuous, configural	0.005	0.003	0.0003	0.0009	0.993
10000, 4, 50/50, continuous, metric	0.004	0.003	0.0003	0.0007	0.997
10000, 4, none, none, none	0.002	0.003	0.0004	0.0008	0.043
10000, 8, 20/80, categorical, configural	0.005	0.004	0.0004	0.0011	1.000
10000, 8, 20/80, categorical, metric	0.005	0.003	0.0004	0.0009	1.000
10000, 8, 20/80, continuous, configural	0.004	0.005	0.0003	0.0005	0.997
10000, 8, 20/80, continuous, metric	0.003	0.005	0.0002	0.0005	0.999
10000, 8, 50/50, categorical, configural	0.003	0.005	0.0002	0.0006	1.000
10000, 8, 50/50, categorical, metric	0.003	0.004	0.0002	0.0004	1.000
10000, 8, 50/50, continuous, configural	0.005	0.003	0.0003	0.0011	0.992
10000, 8, 50/50, continuous, metric	0.004	0.003	0.0003	0.0011	0.997
10000, 8, none, none, none	0.002	0.003	0.0004	0.0008	0.057
400, 4, 20/80, categorical, configural	0.025	0.019	0.0028	0.0014	1.000
400, 4, 20/80, categorical, metric	0.025	0.016	0.0027	0.0013	1.000
400, 4, 20/80, continuous, configural	0.019	0.026	0.0015	0.0032	0.998
400, 4, 20/80, continuous, metric	0.016	0.025	0.0014	0.0032	0.955
400, 4, 50/50, categorical, configural	0.015	0.024	0.0014	0.0020	1.000
400, 4, 50/50, categorical, metric	0.015	0.021	0.0015	0.0018	1.000
400, 4, 50/50, continuous, configural	0.024	0.016	0.0020	0.0021	1.000
400, 4, 50/50, continuous, metric	0.021	0.016	0.0019	0.0019	1.000
400, 4, none, none, none	0.011	0.017	0.0014	0.0065	0.016
400, 8, 20/80, categorical, configural	0.025	0.019	0.0030	0.0014	1.000
400, 8, 20/80, categorical, metric	0.026	0.016	0.0027	0.0014	0.999
400, 8, 20/80, continuous, configural	0.019	0.025	0.0015	0.0031	0.996
400, 8, 20/80, continuous, metric	0.016	0.025	0.0015	0.0031	0.936
400, 8, 50/50, categorical, configural	0.016	0.024	0.0014	0.0019	1.000
400, 8, 50/50, categorical, metric	0.015	0.021	0.0014	0.0018	1.000
400, 8, 50/50, continuous, configural	0.024	0.016	0.0020	0.0020	1.000
400, 8, 50/50, continuous, metric	0.021	0.015	0.0019	0.0021	1.000
400, 8, none, none, none	0.011	0.016	0.0014	0.0040	0.017

Note. SD = standard deviation. Condition: First entry corresponds to sample size, second to number of distractors, third to group size ratio, fourth to type of split covariate, fifth to type of lack of measurement invariance.

Table 10 shows the SRMRs for all conditions in the leaf nodes as well as the corresponding split rates (i.e., power and type I error rate). As can be seen, all SRMRs were < 0.03 with $SD < 0.01$. Differences were most notable between sample sizes, such that SRMRs were smaller with increasing sample size. This seems reasonable as larger samples allow for more accurate model estimation.

4.9 Discussion

We investigated EFA trees as a method to explore and test for MI in a sample of questionnaire data. Our toy examples showed that EFA trees can be used as a simple and straightforward extension of methods that substantive researchers are familiar with. The comprehensive simulation study further highlighted that EFA trees perform well under various conditions. In all conditions, EFA trees demonstrated a high power to detect non-invariance while keeping false-positive splits in the pre-specified range. Ultimately, our goal is to suggest a method that helps researchers to develop questionnaires that took MI into account from the beginning. Additionally, EFA trees can be used as a first tool of exploration when analyzing data before more rigorous steps to test for MI are employed. This is particularly useful when there are no hypotheses about covariates that might cause non-invariance. Even for questionnaires developed as invariant as possible, these tests for MI prior to analyses are indispensable. One should keep in mind here that MI cannot be considered a characteristic of a construct but needs to be addressed for every construct in every study (Vandenberg, 2002).

4.9.1 Why should you use EFA trees?

From a conceptually and theoretically broader perspective, we see three main advantages of EFA trees (and the same applies, in our opinion, to SEM trees and Rasch trees). First, both the seminal review by Vandenberg and Lance (2000) and the more recent one by Putnick and Bornstein (2016) showed that there is a high interest and need for tools that can explore and test for MI. This is good news because addressing MI related issues helps to improve the quality of psychological measurement. In all areas of psychology, improving measurement quality should be a main goal. Otherwise, ever more sophisticated data analysis methods (most notably, machine learning algorithms) cannot unfold their full potential. In fact, as Jacobucci and Grimm (2020) demonstrated, only small amounts of measurement error already diminish the effectiveness of machine learning algorithms to model non-linear effects. Of course, these tree-based methods will not solve all measurement bias related problems. But by equipping researchers with easy-to-use methods whose outputs they are used to interpreting, we can hopefully reduce measurement bias induced by non-invariance or DIF.

Second, EFA trees can assist researchers in shortening questionnaires or in item selection by enabling data-driven exploration of your sample. In practice, one of the main drivers when selecting "good" from "bad" items is the magnitude of factor loadings (Kleka & Soroko, 2018). However, this neglects the fact that even items with a small loading might be important from a content validity standpoint. Even further, there are various reasons why a lack of MI might occur that are arguably more important than loadings when deciding whether to keep or drop/exchange an item. Chen (2008) states many reasons, for example: a) the conceptual meaning or understanding of the construct differs across groups (e.g., for cultural reasons), b) particular items are more applicable for one group than another, c) the item was not translated properly, and/or d) certain groups respond to extreme items differently.

EFA trees do not tell you directly which of these reasons applies to your situation. But they still identify items or whole scales that can then be further explored.[4] In the broadest sense, this might even inform psychological theory development if items are repeatedly shown to be non-invariant between certain groups (Brandmaier & Jacobucci, 2023). Put simply, an item with a small loading might be preferable to an item that works differently between groups (given that the small loading is not due to non-invariance caused by a covariate that was unmeasured and, thus, undetected by an EFA tree).

Third, EFA trees might help to improve the quality of decisions in single-case assessment. In general research, a lack of MI might lead to meaningless results of comparisons between groups. However, in diagnostic decision making on the single-case level, a lack of MI might cause misclassifications. It is common to incorporate diagnostic evidence gathered by tests like personality questionnaires or symptom severity scales when assessing whether a person is suitable for a job or eligible for a certain treatment. Ultimately, besides researching human behavior, this is the main reason why psychological tests are developed in the first place. Thus, it is crucial to

[4]Note that if factor solutions in the nodes are rotated instead of regularized, the items or scales that are identified as non-invariant depend on the exact factor rotation. This is because the solutions are no longer unique and thus different rotations might lead to different interpretations of the solutions. Regularized solutions are unique given a specific type of regularization (e.g., LASSO, ridge, or elastic net) and a specific set of hyperparameters. Changing these settings might again yield different interpretations.

develop questionnaires that are as invariant as possible between all potential target groups (Borsboom, 2006). Of course, this is an overly optimistic goal but we should then at least know for which groups a questionnaire can be used. Imagine using a depression scale that works differently for men and women, such that men receive lower test scores of depressivity even though their true score is equal to that of women. As a consequence, men would on average receive less diagnoses and, thus, less treatment for their depression or women would be overdiagnosed and overtreated in return. Therefore, easy-to-use methods for the assessment of MI on a high level can be a powerful tool to create fair and broadly applicable measures.

4.9.2 How deep is your tree?

One important question we have not yet addressed directly is the depth of EFA trees. We have mostly talked about EFA trees that split the data once but have also shown that deeper trees are possible, revealing interactions between covariates. Theoretically, there is no limit on the depth of a tree (e.g., see Brandmaier et al., 2013b for SEM trees with up to four splits). However, we recommend that you decide on the depth of your tree depending on the goal of your analysis (if multiple interactions between covariates that are associated with non-invariance are present). The main areas of application of EFA trees are the earliest stages of questionnaire development and prior to specific analyses between two groups. In both scenarios, we see two main points to consider when deciding on the depth of your tree: sample size and interpretability.

First, sample sizes in the nodes have to be sufficiently large to allow for stable model estimations. Only then meaningful conclusions about the structure can be drawn. When we consider classic recommendations (Fabrigar et al., 1999) and current practice (Goretzko et al., 2021) regarding sample sizes in EFA, splitting more than once or twice might lead to too few observations in the leaf nodes.

Second, the heterogeneous groups identified by EFA trees should be reasonably interpretable (cf. Zeileis et al., 2008). As mentioned earlier, a split is always dependent on all prior splits. Especially in the earliest stages of questionnaire development, a main goal should be to identify non-invariant groups on a high level. Additionally, as explained in the Introduction, EFA trees use hierarchical clustering. That is, each

split is conditional on the previous split. While this allows to determine the number of heterogeneous groups in a data-driven manner, the allocation of observations to the leaf nodes might not be optimal from a clustering perspective. This is less of a problem with shallow trees, whereas it is amplified when trees become deeper because more interactions are present. Thus, the deeper the tree is grown, we would recommend to be more cautious not to overinterpret the models in the leaf nodes.

4.9.3 Limitations and Future Directions

Inevitably, EFA trees come with a few limitations that researchers should keep in mind when applying the method. One issue when working with a single tree-based algorithm is that it is dependent on the specific sample (Breiman, 2001). To counteract this dependency, ensemble learning methods like *random forests* can be applied. In a random forest, multiple decision trees are grown in parallel and the results of all trees are aggregated into a single, more stable prediction of unseen data. Brandmaier et al. (2016) suggest *SEM forests* as an extension to SEM trees. They argue that SEM forests should not be seen as a "better" version of SEM trees but that both algorithms are complementary analyses. While SEM trees captivate by their interpretability and the information they yield about a sample at hand, specific partitions may not be optimal or may not generalize to new samples. SEM forests, in turn, can be used to obtain more stable estimates about covariates that predict difference in data patterns. Analogously, EFA trees can be extended to *EFA forests*. We want to point out two cautionary notes regarding this extension. First, it should be noted that growing even a single tree can be very expensive from a computational point of view. If a continuous covariate is identified as a split variable, the exhaustive search of order $O(N)$ can take well over one hour to yield a split point (on a standard local machine). Considering typical ensemble sizes of random forest (say 500 single trees), this can be time consuming even with parallelization on two or four cores. Of course, researchers who have supercomputing clusters available can make use of more cores for larger parallelization setups. Second, while the dependence on a specific sample makes decision trees unstable in their predictions of new data, the assessment of MI with respect to the present sample is the primary goal of EFA trees. The main strength of EFA trees lays in interpretability which we regard higher than predictive performance in this context

(cf. Zeileis et al., 2008). Although an ensemble approach like random forests increases the generalizability of predictions, it impedes the inspection of a specific partition. If the goal is to obtain an interpretable structure for a sample at hand, a single EFA tree should be preferred.

As mentioned earlier, EFA models in the tree are estimated using maximum likelihood estimation (MLE). Unfortunately, so far no other estimation method can be applied because the hypothesis tests used to test for parameter differences need a well-defined likelihood (Hothorn et al., 2006; Zeileis & Hornik, 2007; Zeileis et al., 2008). Even though MLE is one of the most commonly used estimation methods for EFA, it is only suitable for multivariate normal data (Fabrigar et al., 1999; Goretzko et al., 2021). With the typical use of Likert-type items in psychological questionnaires (especially when answer options are few), this assumption of normality is questionable. Researchers should evaluate whether MLE is suitable for their data before applying EFA trees. Additionally, future studies are needed to assess the performance of EFA trees under non-normal data, for example with a dichotomous item format.

Another limitation one should keep in mind is that the sensitivity of the tree can only be governed by the level of significance that is set for the hypothesis tests rather than by considering effect sizes. That is, EFA trees are calibrated in a frequentist manner without really taking into account the impact of non-invariance on the subsequent analyses. Measures exist that directly link the degree of non-invariance to the impact it has on substantive analyses between groups (e.g., *EPC-interest*, Oberski, 2014). Moreover, Chen (2007) comprehensively evaluated the sensitivity of common goodness-of-fit indexes like SRMR to lack of MI. However, when using EFA trees, one can calibrate the trees only abstractly by adjusting the level of significance. That is, the higher the level of significance, the higher the sensitivity to detect smaller degrees of non-invariance. Similarly, if sample sizes become larger, smaller degrees of non-invariance become statistically significant without being practically relevant. It is crucial to thoroughly investigate the models in the leaf nodes to identify whether a split is actually meaningful. Here, too, can domain expertise help to identify possible false-positive splits. Future research should investigate measures that could govern the sensitivity of the tree by considering minimum non-invariance thresholds (i.e., a

minimum degree of non-invariance that is deemed relevant for splitting).

The last limitation was raised by Strobl et al. (2015) in the context of Rasch trees and equally applies to EFA trees: If a covariate that causes non-invariance has not been measured, it cannot be detected by the tree. However, if a covariate that is correlated with the relevant missing one is available, non-invariance may still be detected (Strobl et al., 2015). For this reason, a covariate identified for splitting the data cannot simply be interpreted as the root cause of the lack of MI. That is, any split covariate might well be just the observed version of a latent variable causing non-invariance. This again highlights the importance of thoroughly investigating the data and to use EFA trees as a means of exploration.

4.10 Conclusion

EFA trees offer an easy-to-use and well-known approach to exploring data and testing for MI. They are especially useful in areas like personality or clinical psychology where constructs can be multidimensional and complex. We hope to motivate researchers to test for MI in the earliest stages of questionnaire development but also before substantive group comparisons. In this, measurement bias in general research will hopefully be reduced and diagnostic decisions might even become fairer. When it comes down to it, there is hardly any area of psychology or any research question that would not benefit from more measurement invariance. Or, to put it in the words of Meredith (1993) (p. 540): "It should be obvious that measurement invariance [...] are idealizations. They are, however, enormously useful idealizations in their application to psychological theory building and evaluation."

4.11 References

Ammerman, B. A., Jacobucci, R., & McCloskey, M. S. (2019). Reconsidering important outcomes of the nonsuicidal self-injury disorder diagnostic criterion A. *Journal of Clinical Psychology*, *75*(6), 1084–1097. https://doi.org/10.1002/jclp.22754

Asparouhov, T., & Muthén, B. O. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

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

Aust, F., & Barth, M. (2020). *Papaja: Create APA manuscripts with R Markdown.*

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*(11), 176–181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Brandmaier, A. M., Driver, C. C., & Voelkle, M. C. (2018). Recursive partitioning in continuous time analysis. In K. van Montfort, J. H. L. Oud, & M. C. Voelkle (Eds.), *Continuous time modeling in the behavioral and related sciences* (pp. 259–282). Springer.

Brandmaier, A. M., & Jacobucci, R. (2023). Machine-learning approaches to structural equation modeling. In R. A. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 722–739). Guilford Press.

Brandmaier, A. M., Prindle, J. J., McArdle, J. J., & Lindenberger, U. (2016). Theory-guided exploration with structural equation model forests. *Psychological Methods*, *21*(4), 566–582. https://doi.org/10.1037/met0000090

Brandmaier, A. M., Ram, N., Wagner, G. G., & Gerstorf, D. (2017). Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates. *Developmental Psychology*, *53*(5), 996–1012. https://doi.org/10.1037/dev0000274

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013a). Exploratory data mining with structural equation model trees. In J. J. McArdle

& G. Ritschard (Eds.), *Contemporary Issues in Exploratory Data Mining in the Behavioral Sciences* (pp. 96–127). Routledge.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013b). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

de Mooij, S. M., Henson, R. N., Waldorp, L. J., & Kievit, R. A. (2018). Age differentiation within gray matter, white matter, and between memory and white matter in an adult life span cohort. *Journal of Neuroscience*, *38*(25), 5826–5836. https://doi.org/10.1523/JNEUROSCI.1627-17.2018

De Roover, K. (2021). Finding clusters of groups with measurement invariance: Unraveling intercept non-invariance with mixture multigroup factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., & Vermunt, J. K. (2019). On the exploratory road to unraveling factor loading non-invariance: A new multigroup rotation approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 905–923. https://doi.org/10.1080/10705511.2019.1590778

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups.

Psychological Methods, *27*, 281–306. https://doi.org/10.1037/met0000355

Debelak, R., & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement*, *79*(2), 385–398. https://doi.org/10.1177/0013164418777784

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing measurement invariance in the target rotated multigroup exploratory factor model. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 295–314. https://doi.org/10.1080/10705510902751416

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., ... Hothorn, M. T. (2021). Package 'mvtnorm.' *Journal of Computational and Graphical Statistics*, *11*, 950–971.

Goretzko, D., & Bühner, M. (2022). Note: Machine learning modeling and optimization techniques in psychological assessment. *Psychological Test and Assessment Modeling*, *64*(1), 3–21.

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Hirose, K., & Yamamoto, M. (2014). Estimation of an oblique structure via penalized likelihood factor analysis. *Computational Statistics & Data Analysis*, *79*, 120–132. https://doi.org/10.1016/j.csda.2014.05.011

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical*

Statistics, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905–3909.

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Jacobucci, R., Grimm, K. J., Brandmaier, A. M., Serang, S., Kievit, R. A., Scharf, F., … Jacobucci, M. R. (2016). Package 'regsem.'

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 270–282. https://doi.org/10.1080/10705511.2016.1250637

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings, *38*(4), 593–604. https://doi.org/10.1007/BF02291497

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482. https://doi.org/10.1007/BF02289658

Kiers, H. A. L. (1994). Simplimax: Oblique rotation to an optimal target with simple structure. *Psychometrika*, *59*(4), 567–579. https://doi.org/10.1007/BF02294392

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kleka, P., & Soroko, E. (2018). How to avoid the sins of questionnaires abridgement? *Survey Research Methods*, *12*(2), 147–160. https://doi.org/10.18148/srm/2018.v12i2.7224

Komboz, B., Strobl, C., & Zeileis, A. (2018). Tree-based global model tests for polytomous Rasch models. *Educational and Psychological Measurement*, *78*(1), 128–166. https://doi.org/10.1177/0013164416664394

Li, X., Jacobucci, R., & Ammerman, B. A. (2021). Tutorial on the Use of the regsem Package in R. *Psych*, *3*(4), 579–593. https://doi.org/10.3390/psych3040038

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*(4), 361–388. https://doi.org/10.1177/1094428104268027

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

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584. https://doi.org/10.1007/s11336-013-9376-7

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82. https://doi.org/10.1007/s11336-012-9302-4

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–130. https://doi.org/10.21500/20112084.857

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge.

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60. https://doi.org/10.1093/pan/mpt014

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and

reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

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

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347–363. https://doi.org/10.1177/0734282911406661

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research*, *45*(1), 73–103. https://doi.org/10.1080/00273170903504810

Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(4), 576–590. https://doi.org/10.1080/10705511.2018.1558060

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2021). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01689-0

Simpson-Kent, I. L., Fuhrmann, D., Bathelt, J., Achterberg, J., Borgeest, G. S., & Kievit, R. A. (2020). Neurocognitive reorganization between crystallized intelligence, fluid intelligence and white matter microstructure in two age-heterogeneous developmental cohorts. *Developmental Cognitive Neuroscience*, *41*, 100743. https://doi.org/10.1016/j.dcn.2019.100743

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316.

https://doi.org/10.1007/s11336-013-9388-3

Trendafilov, N. T. (2014). From simple structure to sparse components: A review. *Computational Statistics*, *29*(3), 431–454.

Usami, S., Hayes, T., & McArdle, J. (2017). Fitting structural equation model trees and latent growth curve mixture models in longitudinal designs: The influence of model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 585–598. https://doi.org/10.1080/10705511.2016.1266267

Usami, S., Jacobucci, R., & Hayes, T. (2019). The performance of latent growth curve model-based structural equation model trees to uncover population heterogeneity in growth trajectories. *Computational Statistics*, *34*(1), 1–22. https://doi.org/10.1007/s00180-018-0815-x

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*(2), 139–158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*,

67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

5 Study 2: New Developments in Measurement Invariance Testing: An Overview and Comparison of EFA-based Approaches

Sterner, P., De Roover, K., & Goretzko, D. (2024). New Developments in Measurement Invariance Testing: An Overview and Comparison of EFA-based Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *32*(1), 117–135. https://doi.org/10.1080/10705511.2024.2393647

The authors made the following contributions. Philipp Sterner: Conceptualization, Formal Analysis, Methodology, Writing - Original Draft Preparation, Writing - Review & Editing; Kim De Roover: Conceptualization, Resources, Writing - Review & Editing; David Goretzko: Conceptualization, Methodology, Supervision, Writing - Review & Editing.

5.1 Abstract

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

5.2 Introduction

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

Multi-group Confirmatory Factor Analysis (MG-CFA) was originally introduced to test whether a measurement model is invariant across a defined set of groups. However, MG-CFA reaches its limits when many groups have to be compared (e.g., a covariate *nation* with 48 groups; Kuppens et al., 2006). The chance of false-positive findings of non-invariance increases with the number of groups due to multiple testing (Rutkowski & Svetina, 2014). Additionally, this amount of hypothesis tests can make it difficult to tell invariant from non-invariant parameters (Byrne & Vijver, 2010; De Roover et al., 2022). To improve investigations of MI for cases with many groups, more advanced methods have been developed. Raykov et al. (2013) developed a multiple testing procedure to investigate MI that uses the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995). This controls the false-discovery rate rather than the family-wise error rate, resulting in a higher power compared to simple multiple testing of MG-CFAs with Bonferroni correction. Kim et al. (2017) provide a comprehensive overview of

methods to investigate MI with many groups, for example, Multilevel Factor Mixture Modeling and Alignment Optimization (Asparouhov & Muthén, 2014). The majority of these methods developed so far and all methods detailed in Kim et al. (2017) are based on CFA. While the use of CFA allows to incorporate theoretical considerations when investigating MI, it can also be too restrictive in terms of model specification. If the model is slightly misspecified, a CFA-based approach might not accurately recover the true structure of a model (Nájera et al., 2023).

In recent years, new methods have been developed that are based on *exploratory factor analysis* (EFA). As extensions of *multi-group EFA* (MG-EFA; Dolan et al., 2009), researchers developed *mixture multi-group EFA* (MMG-FA; De Roover et al., 2022), *multi-group exploratory factor alignment*[5] (AESEM; Asparouhov & Muthén, 2023), and *EFA trees* (Sterner & Goretzko, 2023). Investigating MI on the basis of EFA avoids the problem of having to assume a (potentially too) restrictive model across all groups. The goal of this paper is to give an overview of these recent developments and to demonstrate the application of EFA-based MI methods. Additionally, because all of these methods inherit the challenge of rotational indeterminacy of the EFA model, we illustrate how these methods can be combined with *multi-group factor rotation* (MGFR; De Roover & Vermunt, 2019). MGFR resolves the rotational indeterminacy per group and locates non-invariant factor loadings by means of hypothesis testing. We demonstrate all of this on an empirical data example from moral psychology (Bago et al., 2022). Because the methods differ in their assumptions and outcomes, a direct comparison does not make too much sense. Instead, we provide a guide on when to use which method in which way. By this, we hope to help researchers to navigate through the extensive literature on EFA-based methods to investigate MI and increase the prevalence of MI testing in social scientific research (Leitgöb et al., 2023; Maassen et al., 2023). To facilitate the application of the presented methods, we provide openly available *R*, *Mplus*, and *Latent Gold* code.

The remainder of the paper is structured as follows: Section 1 outlines the differences between CFA- and EFA-based tests of MI. Section 2 presents the four EFA-based

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

methods (MG-EFA, MMG-FA, AESEM, EFA trees) in detail. For each method, we address the model specification and recommendations on when to use the method, detailing their strengths and weaknesses. Section 3 presents MGFR and an overview table summarizing all methods. Section 4 demonstrates the application of the methods in combination with MGFR. Section 5 discusses differences to and possible combinations with CFA-based methods.

5.3 CFA vs. EFA in MI Testing

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

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

5.4 Multi-group EFA

5.4.1 Model Specification

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

$$\boldsymbol{x}_{i_g} = \boldsymbol{\tau_g} + \boldsymbol{\Lambda_g}\boldsymbol{\xi}_{i_g} + \boldsymbol{\epsilon}_{i_g} \tag{7}$$

where $\boldsymbol{\tau_g}$ is a p-dimensional vector of group-specific intercepts, $\boldsymbol{\Lambda_g}$ is a $p \times m$ matrix of group-specific factor loadings, $\boldsymbol{\xi}_{i_g}$ is a m-dimensional vector of latent factor scores, and $\boldsymbol{\epsilon}_{i_g}$ is a p-dimensional vector of error terms. For maximum-likelihood estimation, the latent factor scores are assumed to be multivariate-normally distributed; specifically, $\boldsymbol{\xi}_{i_g} \sim MVN(\boldsymbol{\alpha_g}, \boldsymbol{\Phi_g})$, where $\boldsymbol{\alpha_g}$ denotes the factor means of group g and $\boldsymbol{\Phi_g}$ the factor (co-)variances. In MG-EFA, the factors are rotationally indeterminate per group, which means there are infinitely many sets of factor solutions which have the same fit to the data but lead to different interpretations of the solution. This has to be resolved per group by a rotation criterion (De Roover & Vermunt, 2019), which often improves interpretability by pursuing simple structure – where each observed variable has near-zero loadings for all factors but one. As already mentioned, we will employ MGFR to address this issue (details will follow in a later section), which not only strives for simple structure but also maximizes the similarity of the rotated loadings across groups. The error terms are also assumed to be multivariate-normal and independent of the factor scores; specifically, $\boldsymbol{\epsilon}_{i_g} \sim MVN(0, \boldsymbol{\Psi_g})$, where $\boldsymbol{\Psi_g}$ is a $p \times p$ diagonal matrix which contains the unique variances of the observed variables in group g. Combining all of the above, we arrive at the group-specific model-implied covariance matrix $\boldsymbol{\Sigma_g} = \boldsymbol{\Lambda_g}\boldsymbol{\Phi_g}\boldsymbol{\Lambda_g}^\top + \boldsymbol{\Psi_g}$.

5.4.2 Testing Procedure

In a factor-analytic context, MI is tested by fitting and comparing increasingly constrained models (De Roover et al., 2022; Vandenberg & Lance, 2000). Many comprehensive guides for MI testing in this context exist, so we will keep this section rather short (see e.g., Putnick & Bornstein, 2016; Van de Schoot et al., 2012). The first step is to test whether configural MI holds. Configural MI means that the construct architecture, that is, the number of latent factors and the location of zero-loadings are equivalent across groups. This is tested by estimating the baseline model in equation (7) per group. Because there are no loadings constrained to zero in MG-EFA, the only model misspecification that could cause the overall model fit to be bad is a different number of latent factors. For example, the baseline models for all groups are estimated with three latent factors but in one group there are actually four latent factors. To partially identify the model, the factor means $\boldsymbol{\alpha_g}$ are set to 0 and the factor covariance matrix $\boldsymbol{\Phi_g}$ is set to an $m \times m$ identity matrix, that is, with factor variances of 1 and factor covariances of 0 (Van de Schoot et al., 2012). In the next step, the invariance of factor loadings, called weak or metric MI, is tested. For this, the fit of the baseline model is compared to the fit of a model in which loadings are constrained to be equal across groups (i.e., $\boldsymbol{\Lambda_1} = ... = \boldsymbol{\Lambda_G}$). If metric MI is supported, latent covariances or relations (e.g., how extraversion relates to other latent variables) can be compared between groups (De Roover et al., 2022). Strong or scalar MI is assessed by comparing the fit of the metric model with the fit of a model with constrained intercepts (i.e., $\boldsymbol{\tau_1} = ... = \boldsymbol{\tau_G}$). If scalar MI is supported, comparisons of latent factor means are warranted (e.g., the means of extraversion). In the last step, strict or residual MI is tested by constraining the unique variances of the observed variables, that is, the diagonal of $\boldsymbol{\Psi_g}$, to be equal across groups. If residual MI holds and factor variances are equal as well, this means that the item reliabilities are equal across groups (e.g., extraversion is measured with the same precision in different groups) (Vandenberg & Lance, 2000). However, this level of MI can be difficult to achieve and is not a prerequisite for the comparison of latent factor means (Chen, 2007; Vandenberg, 2002).

A decrease in fit when estimating a more restricted model is an indication that the tested level of MI is not supported (Chen, 2007; Cheung & Rensvold, 2002); for example,

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

5.4.3 When To Use MG-EFA

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

MG-EFA comes with a lack of flexibility and strong assumptions that have to be made. MG-EFA can only test MI on covariates that are measured and for which hypotheses about non-invariance exist. If a covariate associated with non-invariance is not measured or if there are no hypotheses about non-invariant group constellations, MG-EFA reaches its limits. For example, researchers have to choose the covariate *gender* and form hypotheses about non-invariant groups to test MI on this covariate.

However, more often than not the choice of which covariate to test for non-invariance is not straightforward (Sterner et al., 2024). Testing all available covariates with all potential group constellations leads to the already mentioned multiple testing problem. This is emphasized in cases where a covariate encompasses many groups and nearly impossible when a covariate is continuous (e.g., *age*; Putnick & Bornstein, 2016). To summarize, if you want to test MI for a measured categorical covariate with a small number of groups, and if specifying a CFA model might be too strict, MG-EFA is a good option. If not, you might want to resort to one of the methods presented in the following. We will explain how these methods can find unmeasured clusters of groups for which MI holds, investigate MI along a continuous covariate or identify covariates associated with MI without any hypotheses about them.

5.5 Mixture Multi-group EFA

5.5.1 Model Specification

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

readers interested in the model with cluster-specific intercepts to De Roover (2021). The MMG-EFA model with cluster-specific loadings for group g is

$$f(\boldsymbol{X_g};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{gk}(\boldsymbol{X_g};\boldsymbol{\theta_{gk}}) = \sum_{k=1}^{K} \pi_k \prod_{i_g=1}^{N_g} MVN(\boldsymbol{x_{i_g}};\boldsymbol{\mu_g},\boldsymbol{\Sigma_{gk}}) \tag{8}$$

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

It is important to note that the invariance of parameters within each cluster only holds under the assumption that the correct number of clusters K was extracted. If too few clusters are selected, MI may not hold within each cluster. If too many clusters are selected, MI may hold across some of the clusters. This model selection problem is addressed by combining both the *Bayesian Information Criterion* (BIC; Schwarz, 1978) and the *Convex Hull procedure* (Ceulemans & Kiers, 2006; CHull; Ceulemans & Van Mechelen, 2005). The BIC tries to strike a balance between model fit and complexity by adding a penalty for additional free parameters and larger sample sizes. De Roover et al. (2022) and De Roover (2021) recommend to use the number of groups G for the sample size when computing the BIC (instead of the actual sample size) because the clustering operates at the group level. CHull can be seen as a generalization of the scree test (Cattell, 1966), again trying to balance model fit and complexity. This

is similar to the approach suggested by Lorenzo-Seva et al. (2011) to determine the number of factors to be extracted in EFA. We refer readers to De Roover et al. (2022) and De Roover (2021) for more technical details on these two model selection strategies. It is best to run multiple MMG-EFAs with different numbers of clusters and to choose the solution with the lowest BIC and the highest scree-ratio resulting from CHull. In general, when in doubt about how many clusters to extract, it is recommended to investigate the two or three best solutions. Only if the additional clusters show substantive parameter differences, the solution with more clusters should be preferred over the parsimonious solution with less clusters (De Roover, 2021). Additionally, applying MG-EFA per cluster to test whether MI holds within each cluster can be a way to check if the selected number of clusters is plausible.

5.5.2 When To Use MMG-EFA

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

Similarly, finding clusters of groups according to their measurement parameters helps in pinpointing which items are problematic with regard to MI (De Roover, 2021). By comparing the cluster-specific loadings or intercepts, items that are the source of non-invariance can be identified, again with less pairwise comparisons. Another advantage is that the clustering might help to remedy small group sizes. When group sizes are too small to allow for a precise estimation of group-specific parameters, estimating parameters (e.g., loadings) cluster-specifically helps to achieve more reliable

estimates.

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

5.6 EFA Trees

5.6.1 Model Specification

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

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

$$\sum_{i=1}^{N} \pi(Y_i, \hat{\boldsymbol{\theta}}) = 0 \tag{9}$$

whereby

$$\pi(\boldsymbol{Y}, \boldsymbol{\theta}) = \frac{\partial \Pi(\boldsymbol{Y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{10}$$

is the score function of $\Pi(\boldsymbol{Y}, \boldsymbol{\theta})$.

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

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

5.6.2 When To Use EFA trees

The main advantage of EFA trees is that no hypotheses about covariates potentially associated with (non-)invariance are needed. EFA trees automatically test all covariates for non-invariance, as opposed to (M)MG-EFA where grouping covariates have to be specified. In this, they can simultaneously handle categorical and continuous covariates. If one expects non-invariant groups associated with interactions between covariates, these interactions can be detected in two ways (Zeileis et al., 2008): Either the interaction term is added as a potential split covariate into the algorithm; or, to preserve the exploratory spirit of EFA trees, one could allow "deeper" trees, that is, trees that split the data more than once. All splits in a tree are conditional on all prior splits. Suppose an EFA tree splits the data twice on two different covariates *age* and *gender*: Each leaf node (the final node in a tree) can be seen as a group defined by an interaction between these two covariates that lead to this leaf node, for example, women that are older than 30 years.

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

regularization yields a unique solution. However, changing these settings can again alter the conclusions about MI.

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

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

5.7 Multi-group Exploratory Factor Alignment

5.7.1 Model Specification

Alignment aims at enabling a comparison of latent means across groups when full MI is not supported; that is, when there are some small differences in parameters across groups (Asparouhov & Muthén, 2014, 2023). This is done by first estimating a configural model, that is, a model where all parameters are estimated group-specifically (this corresponds to the model in equation (7)). The factor means and variances of these models are set to 0 and 1, respectively, for each group. In a second step, the

alignment step, the factor means and variances of the groups are chosen so that the amount of non-invariance across groups is minimized. This corresponds to minimizing the differences between loadings and intercepts across groups. It is important to note that the factor means and variances are unidentifiable. As a consequence, the alignment does not change the model fit when searching for optimal values of the factor means and variances. To resolve this unidentifiability and to arrive at the optimal (i.e., "most invariant") values, an alignment function F is minimized with respect to the factor means and variances (Asparouhov & Muthén, 2014, 2023):

$$F = \sum_{m} \sum_{p} \sum_{g_k < g_l} w_{g_k,g_l} f(\lambda_{mpg_k} - \lambda_{mpg_l}) + \sum_{p} \sum_{g_k < g_l} w_{g_k,g_l} f(\tau_{pg_k} - \tau_{pg_l}) \tag{11}$$

where g_k and g_l represent groups k and l (with $k \neq l$) for every possible pair, and λ_{mpg_k} and λ_{mpg_l} (τ_{pg_k} and τ_{pg_l}) indicate the factor loadings (intercepts) of groups k and l, respectively. Because in AESEM cross-loadings are considered, all factors are aligned at the same time, in contrast to the original alignment method where all factors are aligned separately (Asparouhov & Muthén, 2023). w_{g_k,g_l} is a weight that depends on the group sizes, $w_{g_k,g_l} = \sqrt{N_{g_k} N_{g_l}}$, and thus expresses the certainty with which parameters for a group are estimated. The component loss function f is used to scale the observed parameter differences among the groups. It is chosen to be $\sqrt{\sqrt{x^2 + \epsilon}}$, where ϵ is a small number, e.g., 0.001. This function is approximately the same as $\sqrt{|x|}$ with ϵ being added to ensure continuous differentiability (Asparouhov & Muthén, 2014; Robitzsch, 2023). Equation (11) is minimized when the majority of loadings and intercepts are invariant, and only a small number of parameters are (largely) non-invariant (i.e., the number of non-invariant parameters is minimized). Medium-sized non-invariant parameters are avoided by this specific loss function (Kim et al., 2017).

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

parameter (Asparouhov & Muthén, 2014). This coefficient indicates the degree to which a parameter is invariant across groups, ranging from 0 (completely non-invariant) to 1 (completely invariant). The combination of these hypothesis tests and effect size estimates is an indication for the degree of (non-)invariance of a parameter across groups (Flake & McCoach, 2018).

The alignment approach has been extended to the EFA model in Asparouhov and Muthén (2023). The only difference to the procedure just described is that the (unrotated) configural model is rotated first, before being aligned. As usual, the rotation is done by minimizing a rotation criterion (e.g., geomin). Quite naturally, these separate steps of rotation and alignment can also be combined by adding the rotation function to the alignment loss function in equation (11). This joint function is then minimized with respect to the factor means, factor variances, and the rotation criterion (i.e., usually a criterion aiming at simple structure solutions). In order to preserve the order of first rotating and then aligning the model, Asparouhov and Muthén (2023) assign an infinitely large weight to the rotation part of the joint function. As a consequence, the method first estimates a rotated configural model which is then aligned, conditional on the rotated solution.

5.7.2 When To Use AESEM

As already mentioned, AESEM —or alignment, in general— is not a test of MI but enables a comparison of latent means without having to make the assumption of exact MI. Especially in cases with many groups, there are many possibilities of MI being violated, so assuming exact MI is often unrealistic (Davidov et al., 2014). One assumption that has to be made for AESEM, however, is that most measurement parameters are invariant and only few parameters are non-invariant. A rough rule-of-thumb in the literature is that 25% of the parameters can be non-invariant (Asparouhov & Muthén, 2014; Flake & McCoach, 2018; Luong & Flake, 2023). If one is willing to make this assumption, AESEM produces a model with a clear interpretation about (non-)invariance. Researchers are provided with approximate latent means which can be used to compare groups even if exact MI is not supported (Asparouhov & Muthén, 2014, 2023). This makes AESEM a powerful follow-up method for the other methods presented here and elsewhere because it provides a way of handling non-

invariant measurement models. One advantage of alignment in general is that it is well-researched under various conditions and that applications of the method on real data exist. For example, we refer readers to Munck et al. (2018) and Lomazzi (2018) for exemplary applications, to Luong and Flake (2023) for an in-depth tutorial, and to Flake and McCoach (2018) for a simulation study on its performance with polytomous items. Rudnev (2019) details a tutorial on alignment with *Mplus* syntax.

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

5.8 Multi-group Factor Rotation

5.8.1 Model Specification

As we have mentioned several times throughout this paper, EFA models are only determined up to admissible rotations. That is, the estimated solutions can be rotated in infinitely many ways without altering the goodness of fit of the model. To resolve this rotational indeterminacy, a rotation criterion has to be specified (see Browne, 2001 for an overview). What changes with different rotations, however, is the interpretation of the solutions. Depending on the rotation, the loading patterns (i.e., the size and

probably also the allocation of primary and cross-loadings to latent factors) may completely change for a group. This, in turn, affects the conclusions regarding (non-)invariance across groups. Consequently, the choice of the rotation criterion is critical in a multi-group context (De Roover & Vermunt, 2019). If we are only interested in whether all loadings are invariant, the rotation of the solution is irrelevant. This is because a fully invariant factor model will show invariant loading patterns among all groups for every admissible rotation. We would then just compare the fit of the configural model and the metric MI model. If the loadings are invariant, we can impose equal loadings across groups and apply simple structure rotation or target rotation to this single set of loadings. If, however, loadings are non-invariant across group, we need to stick with the group-specific loadings and our goal would be to identify which loadings are non-invariant. This is needed to consider partial MI or item selection, or to reason about potential sources of non-invariance (De Roover & Vermunt, 2019). Solely applying simple structure rotation per group would not be optimal because it does not pursue agreement of the rotated factor loadings between groups. De Roover and Vermunt (2019) introduced MGFR to solve this rotation issue and provide a way of identifying loading differences between groups by means of hypothesis testing.[6] By applying MGFR, the solutions are rotated both to simple structure per group and to agreement between groups. For this, MGFR minimizes a rotation criterion and an agreement criterion (i.e., minimizes disagreement between groups) in a combined multi-group criterion:

$$R^{MG}(\boldsymbol{\Lambda_1}, ..., \boldsymbol{\Lambda_G}) = wR^A + (1 - w)\sum_{g=1}^{G} R_g^{SS} \tag{12}$$

where R^A is an agreement criterion for all groups, R_g^{SS} a simple structure rotation criterion for group g, and $w \in [0, 1]$ a weight to assign relative importance to these two criteria. When optimizing this combined multi-group rotation criterion (by means of constrained maximum likelihood estimation), the group-specific factor variances and covariances are allowed to differ across groups, which helps to unravel differences in loadings from differences in factor (co-)variances (De Roover & Vermunt, 2019). In this,

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

MGFR is similar to AESEM in the sense that it rotates and rescales the parameters. However, as shown in equation (12), it performs these two steps at the same time —so that both rotation and rescaling optimize the agreement— while not considering the item intercepts. This makes MGFR the better alternative when the focus lies on investigating metric MI.

R_g^{SS} can currently be oblimin or geomin, or a target rotation toward an assumed measurement model.[7] For R^A, De Roover and Vermunt (2019) present two possible choices, namely, *generalized procrustes* (GP; Ten Berge, 1977) and *loading alignment* (LA). GP minimizes large loading differences between groups while allowing small differences. This is achieved by applying the least squares principle:

$$R_{GP}^A = \sum_{g_k=1}^{G} \sum_{g_l=g_k+1}^{G} \sum_p \sum_m (\lambda_{g_k pm} - \lambda_{g_l pm})^2 \tag{13}$$

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

LA is closely related to the alignment function in equation (11) but considers only the loadings:

$$R_{LA}^A = \sum_{g_k=1}^{G} \sum_{g_l=g_k+1}^{G} \sum_p \sum_m \sqrt{\sqrt{(\lambda_{g_k pm} - \lambda_{g_l pm})^2 + \epsilon}} \tag{14}$$

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

All EFA-based MI-methods inherit the challenge of resolving rotational indeterminacy in the multi-group case. In the following empirical demonstration of the

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

methods, we thus show how they can be combined with MGFR to achieve interpretable and comparable factor solutions. Table 11 provides an overview of the assumptions, hyperparameters, and capabilities of the presented EFA-based MI methods.

5.9 Empirical Demonstration

5.9.1 Data

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

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

The data further contain many covariates for which (a violation of) MI of the OUS can be investigated. We did not consider every covariate with every method. Rather,

Table 11

Study 2: Overview of methods based on exploratory factor analysis

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

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

we selected for each method the covariate(s) that we think best demonstrate(s) the main advantages of the respective method. In general, we applied all methods simply for didactic purposes to showcase their exemplary application. Our recommendation is *not* to always apply all methods. We looked at the following covariates:

- level of religiosity: continuous on a scale from 1 (lowest) to 10 (highest); $M = 4.21$ $SD = 2.79$
- region: categorical with three levels "Southern" ($N = 4692$), "Eastern" ($N = 2762$), and "Western" ($N = 14292$)
- age: continuous; $M = 26.05$ $SD = 10.25$
- gender: categorical with four levels "male" ($N = 6300$), "female" ($N = 15189$), "other" ($N = 63$), and "I wish not to answer" ($N = 194$)
- country of origin: categorical with 45 levels.

5.9.2 Software

The analyses were run in *R* (version 4.3.1; R Core Team, 2021), *Mplus* (version 8.9), and *Latent Gold* (Vermunt & Magidson, 2016), depending on which method is available in the respective software (see also Table 11). For analyses in R, we used the packages *lavaan* (Rosseel, 2012), *semTools* (Jorgensen et al., 2022), *partykit* (Hothorn & Zeileis, 2015), *mixmgfa* (available at https://github.com/KimDeRoover/mixmgfa/). Additionally, we created the R package *EFAtree* (https://github.com/philippsterner/EFAtree) which implements the EFA trees presented by Sterner and Goretzko (2023). The paper was written using the package *papaja* (Aust & Barth, 2020). All code needed to reproduce the analyses is openly available at https://osf.io/n8x5d/.

5.9.3 Results

5.9.3.1 MG-EFA.

To demonstrate the use of MG-EFA, we investigated MI of the OUS on the covariate region, that is, between eastern, southern, and western participants. Table 12 shows that the configural model (with two latent factors for all groups) has an acceptable model fit. The χ^2-difference tests for both the comparisons of the configural and the metric as well as the metric and the scalar model is significant (both p-values are < 0.005). Judging by these test results, we would have to conclude that neither

loadings nor intercepts are equal across the three groups. However, as mentioned, the χ^2-difference test is highly sensitive to sample size, which is quite large for the data set at hand. Differences in fit indices reveal that the fit of the metric model (where loadings are constrained to be equal across groups) is not worse than the fit of the configural model (ΔRMSEA $= -0.01$, ΔCFI $= 0.00$). Based on cut-off criteria for a comparison of a small number of groups (in our case: three), the conclusion that metric MI is supported seems more suitable (Chen, 2007; Cheung & Rensvold, 2002). When additionally constraining intercepts to be equal across groups, the fit becomes worse (ΔRMSEA $= 0.01$, ΔCFI $= -0.04$). Consequently, scalar MI seems to not be supported. We could conclude that latent covariances or relations (e.g., the correlation between IB and IH) can be compared between the three groups. Latent factor means, on the other hand, should not be compared without additional considerations (e.g., before establishing partial scalar MI).

Table 12
Study 2: Results of multi-group exploratory factor analysis between regions.

Model	χ^2	df	p-value	RMSEA	Δ RMSEA	CFI	Δ CFI
configural	1,185.15	57	0.000	0.052	0.000	0.958	0.000
metric	1,339.90	85	0.000	0.045	-0.007	0.953	-0.005
scalar	2,343.10	99	0.000	0.056	0.011	0.916	-0.037

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

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

loadings on which factor significantly differ across groups. Of course, Wald hypothesis tests to identify significant differences in loadings could also be applied with any other rotation method.

Table 13

Study 2: Unstandardized loading matrices of multi-group exploratory factor analysis of the Oxford Utilitarianism Scale with region as grouping covariate

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

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

Table 13 shows the resulting loading matrices. Table 14 shows the results of Wald hypothesis tests of loading invariance across the three regions. Due to multiple testing, we corrected the p-values with the Benjamini-Hochberg correction to control the false-discovery rate (Benjamini & Hochberg, 1995), using a level of significance of 0.05 (in the table, the corrected p-values are reported). Many main- and cross-loadings are significantly non-invariant across the three regions. For items 1, 5, 6, and 8 on factor IH and item 1, 2, 3, 5, 6, 8, and 9 on factor IB, the null hypothesis of loading invariance is supported by the data. However, because of the large sample size, these hypothesis tests have a high power to detect even small (and possibly irrelevant) loading differences. It is thus important to also inspect the loading matrices to pinpoint especially critical items.[8] Most notable are the loading differences between regions on items 7 ("It is just as wrong to fail to help someone as it is to actively harm them yourself.") and 4 ("If the only way to ensure the overall well-being and happiness of

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

the people is through the use of political oppression for a short, limited period, then political oppression should be used."). The main loading of item 7 is lower for the southern region (compared to eastern and western regions), while it also has a higher cross-loading in this group. Similarly, on item 4, both the southern and the western region have a lower main-loading than the eastern region, where the southern region again shows a notable cross-loading of 0.20 on this item. Attempts to increase MI of the OUS between regions could start with these two items.

Table 14
Study 2: Results of Wald hypothesis tests of loading invariance across the three regions after multi-group exploratory factor analysis

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

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

5.9.3.2 MMG-EFA.

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

larger than 200. This led to 33 countries being considered in the analysis.

Table 15

Study 2: Fit statistics for the ten mixture multi-group exploratory factor analyses of the Oxford Utilitarianism Scale

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

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

First, we conducted MMG-EFAs with one to ten clusters. According to both the BIC (with the number of groups as sample size) and the CHull procedure, the suggested number of clusters is six (see Table 15), which is the solution we selected.

Table 16
Study 2: Composition of the clusters for the six-cluster solution of mixture multi-group exploratory factor analysis

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

Table 16 shows the composition of these six clusters. Each country was assigned to the cluster for which its posterior cluster membership probability $\hat{z}_{gk}$ was highest. It should be mentioned that $\hat{z}_{gk}$ can take on any value between zero and one, which allows groups to have high posterior cluster membership probabilities for more than one cluster. In practice, however, classification uncertainty is rare and limited because groups usually contain enough sample size for the model to be quite certain about their classification. In addition to the names of the countries, we added the region (as assigned in Bago et al., 2022) as well as the geographic region (i.e., the continent a country belongs to). The clustering does not seem to follow an obvious structure in terms of regions or continents. However, what can be concluded is that the loadings

within each cluster are invariant (given that the correct number of clusters was selected). Two countries, China and Hungary, have their own cluster, which means that they do not share equivalent loadings with any other country.[9]

Table 17

Study 2: Unstandardized loading matrices of the mixture multi-group exploratory factor analysis of the Oxford Utilitarianism Scale with clusters as grouping covariate

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

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

Table 17 shows the loading matrix of each cluster after a MGFR (i.e., the cluster membership was used as a new grouping covariate). Again, the weight of the GP agreement criterion was set to 0.5. Table 18 shows the results of Wald hypothesis tests of loading invariance across the six clusters for all combinations of items and factors. As can be seen, all main- and cross-loadings but one are significantly non-invariant (after Benjamini-Hochberg correction). Only for item 5 on the factor IH (which is a cross-loading), the null hypothesis of loading invariance is supported by the data. When comparing the loading matrices across clusters, we can see that for some items in some clusters there was a shift in main- and cross-loadings (e.g., item 1 in cluster 3 and 5, or item 4 in cluster 1). Additionally, some items show large cross-loadings in some clusters, whereas there are no cross-loadings on these items in other clusters. For example, item 8 has a large cross-loading in cluster 3 and 6, but no cross-loading in cluster 1 and 2.

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

Table 18
Study 2: Results of Wald hypothesis tests of loading invariance across the six clusters of mixture multi-group exploratory factor analysis

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

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

A few things should be noted here: First, the large sample size leads to a high power of the Wald hypothesis test, rendering even practically irrelevant loading differences between clusters statistically significant. Second, the higher the number of clusters (or groups, in general), the more difficult it becomes for MGFR to rotate the loading matrices to a solution that is interpretable within each cluster but also comparable between clusters. Thus, it might be beneficial to change the rotation or agreement criterion as well as try different weights between these two criteria according to recommendations by De Roover and Vermunt (2019). This might yield results that are easier to interpret. Table 19 shows the loading matrices when the weight of the agreement criterion was set to 0.1 (i.e., putting more emphasis on simple structure rotation). In our case, while some loadings changed in size, the positions of main- and cross-loadings did not change notably.

Table 19

Study 2: Unstandardized loading matrices of the mixture multi-group exploratory factor analysis of the Oxford Utilitarianism Scale with clusters as grouping covariate with more weight on rotation than on agreement

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

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

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

5.9.3.3 EFA Trees.

Because EFA trees can simultaneously evaluate multiple covariates for MI, we investigated MI with regard to level of religiosity, region, gender, and age. We used the following settings: level of significance was set to $\alpha = 0.005$, the maximum tree depth to three (including the first node, i.e., a maximum number of two splits), and the minimum sample size per node to $n = 400$. These settings allow for an interpretable tree but with possible interactions (due to restricted tree depth) and reliable parameter estimates in each node (due to large minimum sample size per node). At the same time, the low level of significance mitigates the risk of finding practically irrelevant but statistically significant non-invariance. Note that we can only "afford" these rather strict settings because of the large sample. EFA trees could also be used with more liberal settings in smaller samples (see Sterner & Goretzko, 2023).

Table 20
Study 2: Hypothesis test result in the parent node of the EFA tree

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

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

Table 20 shows the hypothesis test results in the parent node of the EFA tree. All p-values are below α, so the covariate with the smallest p-value is selected for splitting (in this case region with a p-value of 2.9×10^{-89}). The tree split the data into a group with only eastern observations and a group with both southern and western observations.

Table 21
Study 2: Hypothesis test result in the eastern node of the EFA tree

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

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

Table 22
Study 2: Hypothesis test result in the southern and western node of the EFA tree

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

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

Table 21 and Table 22 show the hypothesis test results in these two resulting nodes, respectively. In both nodes, the EFA tree split the data on the covariate age[10], resulting in four final leaf nodes:

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

Figure 3 illustrates this tree structure with corresponding sample sizes in the leaf nodes. It is very likely that the tree would have continued to split the data, had we allowed deeper trees. However, this would have decreased the interpretability because it might have led to eight leaf nodes. If interpretation of the resulting partitions (i.e., potential three-way interactions between covariates) provides substantial increase in

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

information gained, deeper trees are easily possible (e.g., see Brandmaier et al., 2013 for a SEM tree with eight leaf nodes). We continued our investigation with four leaf nodes.

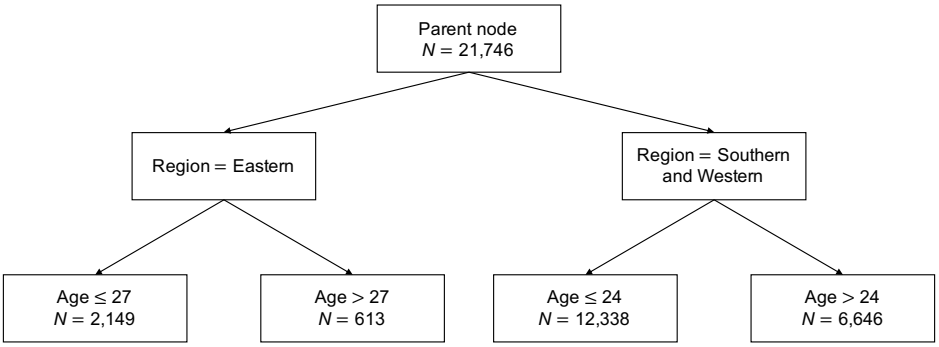

Figure 3
Study 2: Resulting partition after applying EFA trees to the Oxford Utilitarianism Scale data.

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

Table 23
Study 2: Unstandardized loading matrices of the exploratory factor analysis tree of the Oxford Utilitarianism Scale with tree leaf nodes as grouping covariate

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

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

Table 23 shows the loading matrices for all four leaf nodes after applying MGFR (with the weight of the GP agreement criterion set to 0.5). Table 24 shows the results of Wald hypothesis tests of loading invariance across the four leaf nodes for all combinations of items and factors. Again, many main- and cross-loadings are significantly non-invariant (after Benjamini-Hochberg correction). Only for items 1, 5, 6, 8, and 9 on factor IH and item 2, 6, 8, and 9 on factor IB, the null hypothesis of loading invariance is supported by the data. Item 7 ("It is just as wrong to fail to help someone as it is to actively harm them yourself.") sticks out in both "younger" leaf nodes. For younger eastern participants, item 7 shows a lower main-loading compared to both "older" leaf nodes and also has the highest cross-loading in this leaf node. For younger southern-western participants, it has the lowest cross- but also the lowest main-loading (in absolute terms). Item 3 has almost no cross-loading in the younger eastern node but has quite high cross-loadings (> 0.20) in both older nodes (eastern and southern-western). Item 4 is especially noticeable in the younger southern-western node, where its main-loading is more than 0.10 lower compared to the older southern-western node and more than 0.20 lower than in both eastern nodes. Items 7 and 4 here too seem to be the most prominent items with regard to metric non-invariance.

Table 24
Study 2: Results of Wald hypothesis tests of loading invariance across the four leaf nodes of the exploratory factor analysis tree

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

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

5.9.3.4 AESEM.

We used the four nodes generated by the EFA tree (cf. Figure 3) as groups across which AESEM is applied. AESEM could also be applied to a covariate with more groups (e.g., country with its 45 levels). However, by using the results of the EFA trees, we can keep the results easier to interpret while also demonstrating how the methods can be combined.

Table 25
Study 2: Unstandardized loading matrix of exploratory alignment of the Oxford Utilitarianism Scale (weighted average loadings across invariant groups)

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

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

Table 25 shows the average loadings weighted by the sample size across all invariant groups (the detailed Mplus output is available at https://osf.io/n8x5d/). That is, these are the "most invariant" parameters that can be seen as estimates in the groups for which approximate MI holds. This is the case for almost all loadings. On the factor IB, only the loadings of item 7 (which is a main-loading) in the younger southern-western node, and of item 4 (cross-loading) in the younger eastern node were non-invariant compared to all other nodes. On the factor IH, the loadings of item 3 (cross-loading) in the eastern younger node, and item 4 (main-loading) in the younger southern-western node were non-invariant compared to all other nodes. Only for these two groups, these specific parameters cannot be seen as invariant estimates. It can be concluded that the proportion of non-invariant loadings is low; only four out of 72 loadings, that is, 5.6% are non-invariant (four nodes with 18 loadings each = 72 loadings). By applying AESEM to the nodes resulting from an EFA tree, we were thus able to achieve an

approximately invariant set of loadings. This follow-up analysis could even be used to assess the reason why the EFA tree split the data. It might be that the EFA tree split the data because non-invariance on the four mentioned loadings was too large or because of other parameter differences (e.g., factor covariances or residual variances), given that most loadings are approximately non-invariant (after alignment).

5.9.4 Synthesis of the Results

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

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

significantly different, could be taken into account, too.

In closing, we want to highlight again that we applied all methods for didactic purposes. The choice of methods for each individual application depends on the data set, the available covariates, the research question, and the assumptions one is willing to make (cf. Table 11).

5.10 Discussion

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

The main take-away of our presentation and demonstration is that the optimal choice of a method depends on the question you want to answer, combined with the specificities of the data at hand. A detailed (yet not exhaustive) overview of prerequisites and capabilities of each method is given in Table 11. Ideally, the methods are combined to thoroughly scrutinize the data for MI. For example, the clusters or nodes resulting from MMG-EFA or EFA-trees, respectively, can be used as groups in MG-EFA, and the resulting loading matrices can be rotated by MGFR. In this, covariates with many groups can be reduced to a smaller number of clusters or nodes, for which we can then, for example, investigate scalar MI by means of hypothesis testing. We refrained from addressing scalar MI due to the focus of EFA-based methods on metric MI. However, as we will discuss shortly, EFA-based methods could function as a precursor of CFA-based investigations of scalar MI. Even further, clusters or nodes resulting from prior analyses could be used as groups in multi-group (E)SEM (Asparouhov & Muthén, 2009), allowing us to then model structural relations between our constructs of interest.

As mentioned in the introduction, EFA-based methods differ from CFA-based methods mainly in the fact that no (potentially overly restrictive) zero-loadings have to

be imposed. This allows for a more detailed investigation of metric MI because violations of metric MI due to cross-loadings can be considered. However, the investigation of scalar MI is hampered. For example, with EFA trees, the intercepts cannot be included in the model estimation (in *lavaan* language: the argument *meanstructure* must not be set to *TRUE*). The intercepts of the items (the parameters we want to test for invariance) are intertwined with the factor means (the parameters we want to compare between groups). Even if the intercepts were equal across groups, an EFA tree would split the data if the factor means were different between groups. Similarly, the results of MMG-EFA with clustering based on loadings do not consider (non-)invariance of intercepts. It is possible to cluster the groups on both loadings and intercepts. However, this entails the assumption that there is one underlying clustering for both of these sets of parameters (Leitgöb et al., 2023). It is thus more advisable to first cluster the groups based on the loadings and then, per obtained cluster, continue to cluster the groups based on the intercepts. When applying AESEM, both loadings and intercepts are considered and, thus, scalar invariance is also investigated. But because we are also investigating all cross-loadings for invariance, there are more loadings than intercepts that are being estimated and aligned (if the specified model has two or more factors). It should be examined whether this potential dominance of loadings (when minimizing equation (11)) has some undesired effect on the assessment of scalar MI.

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

5.10.1 Future Research

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

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

5.11 References

Asparouhov, T., & Muthén, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Asparouhov, T., & Muthén, B. O. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

Asparouhov, T., & Muthén, B. O. (2023). Multiple Group Alignment for Exploratory and Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(2), 169–191. https://doi.org/10.1080/10705511.2022.2127100

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

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

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification: Two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods & Research*, *36*(1), 48–86.

Brandmaier, A. M., Oertzen, T. von, McArdle, J. J., & Lindenberger, U. (2013).

Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*(1), 111–150.

Byrne, B. M., & Vijver, F. J. R. van de. (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. https://doi.org/10.1080/15305051003637306

Cao, C., & Liang, X. (2022a). Sensitivity of Fit Measures to Lack of Measurement Invariance in Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 248–258. https://doi.org/10.1080/10705511.2021.1975287

Cao, C., & Liang, X. (2022b). The Impact of Model Size on the Sensitivity of Fit Measures in Measurement Invariance Testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 744–754. https://doi.org/10.1080/10705511.2022.2056893

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 133–150. https://doi.org/10.1348/000711005X64817

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

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Test-

ing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement Equivalence in Cross-National Research. *Annual Review of Sociology*, *40*(1), 55–75. https://doi.org/10.1146/annurev-soc-071913-043137

De Roover, K. (2021). Finding Clusters of Groups with Measurement Invariance: Unraveling Intercept Non-Invariance with Mixture Multigroup Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

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

De Roover, K., & Vermunt, J. K. (2019). On the Exploratory Road to Unraveling Factor Loading Non-invariance: A New Multigroup Rotation Approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 905–923. https://doi.org/10.1080/10705511.2019.1590778

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

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

Dolan, C. V., Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing Measurement Invariance in the Target Rotated Multigroup Exploratory Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 295–314. https://doi.org/10.1080/10705510902751416

Flake, J. K., & McCoach, D. B. (2018). An Investigation of the Alignment Method With

Polytomous Indicators Under Conditions of Partial Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 56–70. https://doi.org/10.1080/10705511.2017.1374187

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521.

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research*, *16*(1), 3905–3909.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), 36–44. https://doi.org/10.1038/s41562-019-0734-z

Jennrich, R. I. (1973). Standard errors for obliquely rotated factor loadings. *Psychometrika*, *38*(4), 593–604. https://doi.org/10.1007/BF02291497

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

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

org/package=semTools

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kuppens, P., Ceulemans, E., Timmerman, M. E., Diener, E., & Kim-Prieto, C. (2006). Universal Intracultural and Intercultural Dimensions of the Recalled Frequency of Emotional Experience. *Journal of Cross-Cultural Psychology*, *37*(5), 491–515. https://doi.org/10.1177/0022022106290474

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

Lomazzi, V. (2018). Using Alignment Optimization to Test the Measurement Invariance of Gender Role Attitudes in 59 Countries. *Methods, Data, Analyses : A Journal for Quantitative Methods and Survey Methodology (Mda)*, *12*(1), 77–103. https://doi.org/https://doi.org/10.12758/mda.2017.09

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. IAP.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull Method for Selecting the Number of Common Factors. *Multivariate Behavioral Research*, *46*(2), 340–364. https://doi.org/10.1080/00273171.2011.564527

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

planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

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

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

Marsh, H. W., Morin, A. J. S., Parker, P. D., & Kaur, G. (2014). Exploratory Structural Equation Modeling: An Integration of the Best Features of Exploratory and Confirmatory Factor Analysis. *Annual Review of Clinical Psychology*, *10*(1), 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Application*, *6*(1), 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–130. https://doi.org/10.21500/20112084.857

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

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement Invariance in Comparing Attitudes Toward Immigrants Among Youth Across Europe in 1999 and 2009: The Alignment Method Applied to IEA CIVED and ICCS. *Sociological Methods & Research*, *47*(4), 687–728. https://doi.org/10.1177/0049124117729691

Nájera, P., Abad, F. J., & Sorrel, M. A. (2023). Is exploratory factor analysis always to be preferred? A systematic comparison of factor analytic techniques

throughout the confirmatory–exploratory continuum. *Psychological Methods*, *Advance Online Publication*. https://doi.org/10.1037/met0000579

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

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

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raykov, T., Marcoulides, G. A., & Millsap, R. E. (2013). Factorial Invariance in Multiple Populations: A Multiple Testing Procedure. *Educational and Psychological Measurement*, *73*(4), 713–727. https://doi.org/10.1177/0013164412451978

Robitzsch, A. (2022). *Sirt: Supplementary Item Response Theory Models*. https://cran.r-project.org/web/packages/sirt/index.html

Robitzsch, A. (2023). Implementation Aspects in Invariance Alignment. *Stats*, *6*(4), 1160–1178. https://doi.org/10.3390/stats6040073

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

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

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*, 31–57. https://doi.org/10.1177/0013164413498257

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

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A Review of Measurement Equivalence in Organizational Research: What's Old, What's New, What's Next? *Organizational Research Methods*, *25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Sörbom, D. (1974). A General Method for Studying Differences in Factor Means and Factor Structure Between Groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Sterner, P., & Goretzko, D. (2023). Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30:6*, 871–886. https://doi.org/10.1080/10705511.2023.2188573

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A Causal Framework for the Comparability of Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12. https://doi.org/10.1080/10705511.2024.2339396

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, *42*(2), 267–276. https://doi.org/10.1007/BF02294053

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

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J. (2002). Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research*

Methods, *5*(2), 139–158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Vermunt, J. K., & Magidson, J. (2016). *Technical Guide for Latent GOLD 5.1: Basic, Advanced, and Syntax*. Belmont, MA: Statistical Innovations Inc.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508. https://doi.org/10.1111/j.1467-9574.2007.00371.x

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

5.12 Appendix

Table 26

Study 2: Items and corresponding subscales of the OUS (Kahane et al., 2018)

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

Note. IB = Impartial Beneficence; IH = Instrumental Harm

6 Study 3: A Causal Framework for the Comparability of Latent Variables

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A Causal Framework for the Comparability of Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(5), 747–758. https://doi.org/10.1080/10705511.2024.2339396

The authors made the following contributions. Philipp Sterner: Conceptualization, Methodology, Formal Analysis, Visualization, Writing - Original Draft Preparation, Writing - Review & Editing; Florian Pargent: Conceptualization, Methodology, Writing - Review & Editing; Dominik Deffner: Methodology, Writing - Review & Editing; David Goretzko: Conceptualization, Methodology, Writing - Review & Editing, Supervision.

6.1 Abstract

Measurement invariance (MI) describes the equivalence of measurement models of a construct across groups or time. When comparing latent means, MI is often stated as a prerequisite of meaningful group comparisons. The most common way to investigate MI is multi-group confirmatory factor analysis (MG-CFA). Although numerous guides exist, a recent review showed that MI is rarely investigated in practice. We argue that one reason might be that the results of MG-CFA are uninformative as to why MI does not hold between groups. Consequently, under this framework, it is difficult to regard the study of MI an interesting and constructive step in the modeling process. We show how directed acyclic graphs (DAGs) from the causal inference literature can guide researchers in reasoning about the causes of non-invariance. For this, we first show how DAGs for measurement models can be translated into the path diagrams used in the linear structural equation model (SEM) literature. We then demonstrate how insights gained from this causal perspective can be used to explicitly model encoded causal assumptions with moderated SEMs, allowing for a more enlightening investigation of MI. Ultimately, our goal is to provide a framework in which the investigation of MI is not deemed a "gateway test" that simply licenses further analyses. By enabling researchers to consider MI as an interesting part of the modeling process, we hope to increase the prevalence of investigations of MI altogether.

6.2 Introduction

With increasingly larger and culturally diverse data sets available, social and behavioral scientists are able to research human experiences and behavior in much broader contexts. For example, extensive studies have been conducted on cultural differences in moral judgement (Bago et al., 2022), prosocial behavior (House et al., 2020), and the values of emotions in societies (Bastian et al., 2014). These new opportunities come with new challenges: we need transparent and objective rules about how to adequately compare groups and under which assumptions we are allowed to generalize results from one group to another. Recently, Deffner et al. (2022) have presented a detailed framework based on *causal inference* that does just that: Following simple graphical rules of so-called *directed acyclic graphs* (DAGs), their framework enables researchers to draw inferences and derive licensing assumptions about which comparisons and generalizations are warranted. Researchers working with variables that are observable, like dictator game choices in the examples of Deffner and colleagues, can readily draw on these authors' framework. However, as Deffner et al. (2022) themselves state, psychologists are often interested in the constructs underlying the observed variables (Westfall & Yarkoni, 2016). As psychologists, we do not care whether you reported you enjoy going out with friends — we care about how *extraverted* you are. If we use observed variables as direct representations of the underlying construct (e.g., by building a sum score of questionnaire items), we disregard the *measurement error* inherent in all psychological measures (Lord & Novick, 1968; Van Bork et al., 2022). Ignoring this measurement error in the modeling process can lead to distorted inference. Our model would not be able to distinguish between variation in item responses caused by the construct and variation caused by error (also referred to as unique item variance). Westfall and Yarkoni (2016) for example showed that disregarding measurement error leads to inflated type-I-error rates when trying to statistically control for confounding covariates. As a remedy, they suggest using *structural equation models* (SEM), which are models that explicitly include the measurement error (Bollen, 1989). In a SEM, constructs are modeled in a *measurement model*, where a latent variable and the unique error jointly cause the observed variables (Van Bork et al., 2022). Relationships between constructs are modeled in the *structural model* (Mulaik, 2009). While the use of measurement models allows us to take the measurement error into account,

it poses a new challenge for comparisons between groups. In order to be able to meaningfully compare groups, we have to make sure that any difference between groups occurs only due to true differences (i.e., differences in the latent variable), not due to measurement differences (Meuleman et al., 2022). This characteristic is called *measurement invariance* (MI) and means that the measurement models are equivalent across groups (Meredith, 1993; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Although numerous guides (e.g., Putnick & Bornstein, 2016; Van De Schoot et al., 2012) and methods (e.g., Kim et al., 2017) for investigating MI exist, a recent review showed that it is very rarely done in practice (Maassen et al., 2023). The reasons for this are surely diverse. We argue that one reason might be that researchers currently have only little guidance on how to regard the study of MI an interesting and constructive step in the modeling process. By viewing MI as an informative aspect by itself, we might be able to learn more about psychological constructs. For this, a framework is needed that lets us reason about how and why constructs and measures thereof function differently across groups.

As Deffner et al. (2022) briefly explained, DAGs can be used to depict cases of measurement (non-)invariance. Consequently, DAGs might be a useful tool for reasoning about when latent variables are comparable and generalizable. Our aim is to pick up where Deffner et al. (2022) left off: we want to extend their framework to the case where claims on the construct-level are of interest so that MI is an additional part of the modeling process. The article is structured as follows: First, we briefly introduce the language of DAGs, which are often used in causal inference, and provide a translation to path diagrams for measurement models used in the psychometric SEM literature. Second, we outline the current practice of investigating MI and give a summary of options on how to proceed when MI does not hold. Third, after framing MI as a causal concept, we demonstrate how DAGs can be used to depict non-invariance by encoding assumptions about possible causes of group differences. Fourth, we illustrate in a simulated and an empirical example how following the current practice of investigating MI might miss important aspects of non-invariance. We show how considering the whole causal model instead can help researchers to make more informed modeling choices.

6.3 From DAGs to Measurement Models

We start by clarifying and defining the terms used throughout this paper. As already mentioned, DAGs are graphical objects used in causal inference to depict causal relationships between variables (Elwert, 2013; Pearl, 1998, 2012). They consist of nodes (the variables) which are connected by edges (directed arrows between these nodes). If a variable is unobserved (latent), it we enclose it by a dashed circle. An edge between two variables A and B, denoted by $A \rightarrow B$, means that A has a causal effect on B. DAGs are called *directed* because only single-headed arrows are allowed[11], and *acyclic* because no variable is allowed to be a cause of itself. In general, there are three different causal structures, with which any set of nodes can be described (Deffner et al., 2022; Elwert, 2013; Rohrer, 2018):

- The confounder: $A \leftarrow B \rightarrow C$, that is, the confounder B causes both A and C.
- The chain (psychologists know this as a mediator): $A \rightarrow B \rightarrow C$, that is, A causes C through the mediator B.
- The collider: $A \rightarrow B \leftarrow C$, that is, A and C both cause the collider B.

By following the arrows from one variable to another, we can identify the individual paths by which these variables are connected. For all of these constellations exist clear rules of independences between variables (Mulaik, 2009; Pearl, 2012). We say that two variables are conditionally independent if they are unrelated given a (possibly empty) set of other variables. For the confounder and the chain, conditioning on (also: adjusting for) the variable "in the middle" renders the other two variables independent. In this case, we write $A \perp\!\!\!\perp C \mid B$, meaning that A and C are independent, conditional on B. For the collider, A and C are unconditionally independent; conditioning on B would in turn render them dependent and produce a non-causal association. Thus, conditioning on a variable closes the path (i.e., "stops the flow of information") in the case of confounding and mediating variables but opens a non-causal path (i.e., "allows the flow of information") in the case of colliders (Elwert, 2013). Conditioning can be achieved by including the variable as a predictor in the model but also by specific

[11] Double-headed arrows are sometimes used in DAGs to depict an unobserved common cause between two variables (Elwert, 2013). However, a double-headed arrow between A and B is identical to $A \leftarrow U \rightarrow B$, where U is the unobserved common cause of both A and B. We restrict ourselves to the use of single-headed arrows in this paper.

sampling or experimental designs (Rohrer, 2018). If a path between two variables is closed, the path is said to be *d-separated* (Pearl, 1988). The risk of conditioning on the "wrong" variable or of missing a variable that should be conditioned on highlights that it is crucial to clearly define the causal relationships between variables prior to analyzing or modeling the data. Failure to do so can lead to spurious associations and distorted inference, for example by accidentally opening paths between variables that should remain closed. We refer readers to Rohrer (2018) and Wysocki et al. (2022) for comprehensive guides on how to approach data analysis from a causal inference perspective.

It is important to note that DAGs depict the causal relationships between a set of random variables without imposing particular distributions or functional forms of the relationships (Greenland & Brumback, 2002; Rohrer, 2018; Suzuki et al., 2020). Their strength lies in making assumptions about the relationships between variables explicit and thereby revealing testable implications between them. That is, if the DAG depicts the true data-generating process, applying the graphical rules of (in)dependences tells us which associations should and should not be observable in the data (Elwert, 2013). Even if a DAG does not fully represent the true data-generating process, it would still be useful because all inferences rely on assumptions and a DAG might help to identify the ones that are otherwise made implicitly. If we are not willing to make any assumptions, no analysis can be reasonably justified (Deffner et al., 2022). In this spirit, when setting up a DAG, it is helpful to view the absence of arrows as strong assumptions and their presence as weak ones (Bollen & Pearl, 2013; Elwert, 2013). An omitted arrow between two variables assumes that the direct causal effect is exactly zero, whereas an arrow assumes some form of relationship without specifying its strength or functional form. Thus, the less we are certain about relationships between variables, the more arrows we should draw.

To bridge the gap between DAGs and path diagrams for SEM —and more specifically, measurement models— it is helpful to view DAGs as non-parametric SEMs (Bollen & Pearl, 2013; Pearl, 2012). A non-parametric SEM is a model in which we do not make assumptions about the functional form of the associations between variables. Consider the DAG of a simple measurement model in Figure 4.

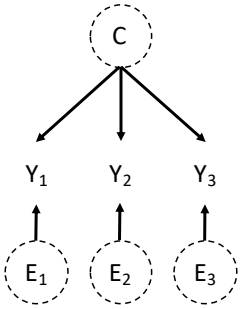

Figure 4
Study 3: Simple DAG of a measurement model where the observed variables Y_1, Y_2, and Y_3 are caused by a latent common factor C and latent unique error terms E_1, E_2, and E_3.

The observed variables Y_1, Y_2, and Y_3 are caused by the unobserved (latent) variables C and E_1, E_2, and E_3. C is called the common factor and interpreted as a common cause of Y_{1-3} (Van Bork et al., 2022). Each Y also has its unique cause E that is independent of C. Interpreting this DAG as a non-parametric SEM, we can formally describe the vector of observed variables $\boldsymbol{Y}$ as $\boldsymbol{Y} = f(C, \boldsymbol{E})$. Typically, when dealing with SEMs, we assume that the relationships are linear and that the variables follow certain distributions. This gives rise to the equation for measurement models in SEM (Mulaik, 2010) [12]:

$$\boldsymbol{Y} = \boldsymbol{\tau} + \boldsymbol{\Lambda} C + \boldsymbol{E} \tag{15}$$

Here, $\boldsymbol{\Lambda}$ is the matrix of path coefficients (called loadings), quantifying the strength of the relationship between the observed variables $\boldsymbol{Y}$ and the latent variable C, $\boldsymbol{\tau}$ is the vector of intercepts of $\boldsymbol{Y}$, and $\boldsymbol{E}$ is the vector of unique error terms of $\boldsymbol{Y}$ which cannot be explained by C. In addition to this structural assumption, the following distributional assumptions are often made for estimation purposes: $C \sim N(\alpha, \Phi)$ and $\boldsymbol{E} \sim MVN(0, \boldsymbol{\Psi})$. α and Φ are the expectation and the variance of C, respectively. The variances of the errors $\boldsymbol{E}$ are captured on the diagonal of $\boldsymbol{\Psi}$ (usually, errors are assumed to be uncorrelated, so the off-diagonal entries of $\boldsymbol{\Psi}$ are 0). The covariance of the data is defined as $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$ (Jöreskog, 1967); that is, variation in the data can be decomposed into a part that is explained by the common factor and an error

[12] Without loss of generality, we are assuming a one-dimensional construct (only one common factor C) and drop the person-indeces i for better readability.

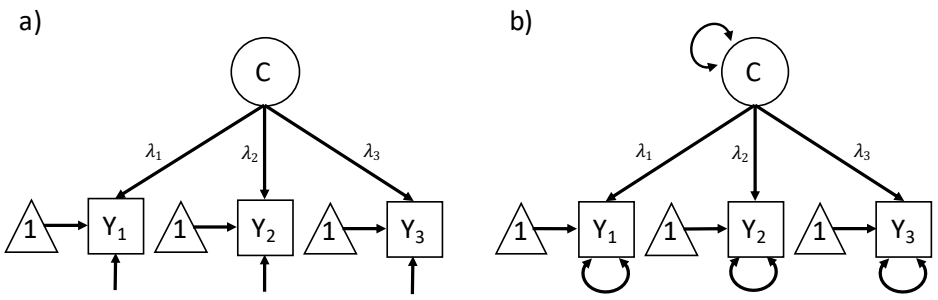

Figure 5
Study 3: Simple path diagram of a measurement model. a) LISREL style: Only error variances are depicted by an arrow without a node pointing into all endogeneous variables (here: the observed variables); b) RAM style: variances of both endogeneous and exogeneous variables are depicted by a double-headed arrow-loop (here: error variances and variances of the latent variables).

part.

The assumption of linearity lets us now translate our measurement model from a DAG (Figure 4) to a path diagram (Figure 5), which is a common form of diagram in the psychological literature (see Epskamp, 2015 for definitions and visualizations of different styles of path diagrams). Latent variables (in our case: C) are enclosed by a circle. Error terms (in our case: E) are not included explicitly. Instead, their variances are depicted by an arrow pointing into its corresponding observed variable (LISREL style, Figure 5a)) or by a double-headed arrow-loop on the observed variable (RAM style, Figure 5b)). Only in RAM style, the variance of the exogeneous variables (in our case: C) are also depicted by a double-headed arrow-loop. The observed variables (in our case: Y) are enclosed by a rectangle, their intercepts are depicted by a triangle. Because we assume that all relationships between variables are linear, we can use path coefficients, that is, a single number on each arrow, to quantify the relationship λ between C and Y.[13] In a DAG, this is not possible because in potentially non-linear relationships the value of the path coefficient between $C \rightarrow Y$ depends on the value of C. When comparing Figure 4 and Figure 5, we can now see that by making structural and distributional assumptions about our causal model, we can translate the DAG of

[13]In path diagrams, double-headed arrows between observed variables (i.e., items) are sometimes used to depict correlated error terms (i.e., item responses that are correlated even after conditioning on the latent variable). This is closely related to the double-headed arrows in DAGs mentioned in an earlier footnote. Correlated errors are equivalent to unobserved confounding, that is, failure to model all influence on the item response besides the latent variable. In the literature on item response models, this is often called local dependence (Kreiner & Christensen, 2011).

our simple measurement model into a path diagram.

The relation between DAGs and path diagrams for SEMs has been shown in the literature (see, e.g., Kunicki et al., 2023 for a comparison) but —to the best of our knowledge— has so far not been extended explicitly to measurement models[14]. We argue that embedding measurement models within wider causal relationships represented by DAGs can help researchers to investigate MI in a more informative manner. In the following, we briefly outline how MI is primarily investigated. Subsequently, we showcase how DAGs can be used to depict (non-)invariance and to decide which variables have to be included in our model. We illustrate how DAGs can be used to investigate assumed causes of non-invariance that might be missed by the current approach.

6.4 Current Practice of Investigating Measurement Invariance

MI is rarely considered in empirical studies on latent variables (Maassen et al., 2023). Specifically, Maassen and colleagues investigated the practice of MI testing for 918 latent mean comparisons in 97 articles in the two journals *PLOS ONE* and *Psychological Science*. They found that references regarding MI in these two influential journals were made for only 40 (4%) of the 918 latent mean comparisons. Additionally, none of these tests could be reproduced due to unavailable data or lack of details in reporting of MI testing procedures. It is thus not clear how many claims about latent variable differences between groups in the literature are actually attributable to true differences and how many occurred due to measurement non-invariance. By no means do we want to imply that researchers who do not consider MI are not rigorous. Rather, our argument is directed against the current practice of investigating MI. As we will outline below, the current approach does not provide much information about the role of (non-)invariance in the data-generating process. Additionally, it does not inform researchers about principled measures to choose an appropriate model to investigate or consider MI in their analyses.

Prevailingly, MI is (in its simplest form) tested by *multi-group confirmatory factor analysis* (MG-CFA) with G groups (Jöreskog, 1971): A covariate that defines the

[14]But see Bollen and Pearl (2013) who briefly touch on measurement models in combination with causality.

groups to be compared is chosen, for example the covariate *Region* with two groups *western* and *eastern*. First, a factor analysis model (see equation (15)) is estimated per group, that is, with group-specific loadings, intercepts, and unique variances. This is called a configural model. A combined goodness-of-fit measure for both groups is calculated, for example the root mean squared error of approximation (RMSEA) or the comparative fit index (CFI). A bad fit of the configural model is an indication that the model itself is misspecified (i.e., missing paths between observed and latent variables or wrong number of latent variables in one or more groups). Next, a second model is estimated but now the loadings are constrained to be equal across groups (i.e., $\mathbf{\Lambda}_g = \mathbf{\Lambda}_k$ for all $g, k \in 1, ..., G$). If the overall fit of this model does not drop compared to the configural model, metric (or weak) MI is supported, that is, loadings are equal across groups. In a third model, in addition to the loadings, the intercepts of the observed variables are constrained to be equal across groups (i.e., $\boldsymbol{\tau}_g = \boldsymbol{\tau}_k$ for all $g, k \in 1, ..., G$). If the overall fit is not worse than the fit of the metric model, scalar (or strong) MI holds. If scalar MI is supported, comparisons of latent means are warranted (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). As a rule-of-thumb, an increase of 0.01 of the RMSEA or a decrease of 0.01 of the CFI when comparing two nested models could be considered a violation of MI (Chen, 2007; Cheung & Rensvold, 2002). Rutkowski and Svetina (2014) propose more liberal values of 0.03 in RMSEA-increase or 0.02 in CFI-decrease when testing for metric MI and when the number of groups is high. Nonetheless, because cut-off values depend on both model complexity and sample size, researchers should not blindly follow these recommendations (Goretzko et al., 2023). Since the models are nested, a stricter comparison by means of a χ^2-difference hypothesis test is possible as well. However, this test is sensitive to sample size, so using fit indices is considered more suitable (De Roover et al., 2022). Beyond scalar MI, residual MI could be tested by comparing the scalar model with a model in which the unique variances are constrained to be equal. Because this level of MI is difficult to achieve and not a prerequisite of latent mean comparisons, it is often not considered.

The results of this investigation do not provide any information on *why* MI is not supported. Thus, it is not obvious what to do if we find that MI does not hold or if we want to consider it as a part of the whole modeling process. We briefly outline a few options on how to proceed in this case. We refer readers to Leitgöb et al. (2023) for a

detailed account of the approaches mentioned below. First, one could aim for partial MI. This is done by identifying so called anchor items, that is, items whose parameters are invariant across groups. By constraining parameters of these anchor items to be equal across groups and and allowing the remaining parameters to differ, partial MI can be established (Vandenberg & Lance, 2000). Unfortunately, there is no clear answer to the question of how many parameters have to be equal across groups to allow for meaningful latent mean comparisons (Putnick & Bornstein, 2016). Additionally, the identification of anchor items is far from trivial (Sass, 2011; Steenkamp & Baumgartner, 1998) and the wrong choice can again bias latent mean comparisons (Belzak & Bauer, 2020; Pohl et al., 2021). Second, more advanced methods to investigate MI could be applied, for example from the literature on differential item functioning (Bauer et al., 2020; Kopf et al., 2015; Strobl et al., 2015; Tutz & Schauberger, 2015) or on SEM (Asparouhov & Muthén, 2014; Brandmaier et al., 2013; De Roover et al., 2022; Schulze & Pohl, 2021; Sterner & Goretzko, 2023). However, all of these methods entail specific assumptions about the variables in the data and the relationships between them. To exploit their full potential, it is crucial to explicitly consider these assumptions in order to make informed modeling decisions. Luong and Flake (2023) provided a detailed example of how taking into account the underlying assumptions of advanced methods to investigate MI could look like. Third, at some point, we might have to accept that MI does not hold (Leitgöb et al., 2023; Rudnev, 2019). This, however, is an important finding by itself and should be the starting point of further exploration (for an example, see Seifert et al., 2024). Especially when constructing or revising psychological tests or questionnaires, thoroughly exploring why a measure functions differently across groups can help us to learn more about the construct itself. As Putnick and Bornstein (2016) put it, investigating MI should not be considered a "gateway test" that licenses us to further analyze our data. Rather, it should be viewed as an integral part of the whole modeling process.

What is, in our opinion, currently missing is a theoretical framework in which a potential lack of MI can be explored. Specifically, a framework is needed which lets us reason about the *causes* of non-invariance. As mentioned, because MI is usually only investigated with regard to the covariate that defines the groups we want to compare, the only information we get is *that* MI is violated. Under this approach, it is

difficult to communicate assumptions about *why* MI does not hold. Researchers can therefore not properly decide how their statistical models to investigate MI should look like. Consequently, they are unable to make full use of the broad arsenal of advanced methods. By outlining the causal foundations of MI, we now demonstrate how DAGs can be used to depict (a lack of) MI and to make informed modeling choices.

6.5 The Causal Foundations of Measurement Invariance

When looking at seminal papers on MI, one could argue that MI was a causal concept from the very beginning. Mellenbergh (1989) depicted non-invariance (he called it item bias) by some form of DAG and speaks of causal influences as well as conditional independencies between observed variables (items), latent variables (traits), and groups. Similarly but more formally, Meredith (1993) would define our observed variable Y as measurement invariant with respect to selection on some other variable V if Y and V are independent, conditional on the latent variable C. Thus, MI is formally defined as

$$f(Y|V, C) = f(Y|C) \tag{16}$$

where $f(\cdot)$ is the density function. That is, conditional on the common factor C, the distribution of the observed variables Y is independent of any variable V ($Y \perp\!\!\!\perp V \mid C$). V is usually assumed to be an observed covariate (e.g., age, region, gender, etc.) but could also be a latent variable. MI thus means that the measurement model is equivalent in any group within the population. Borsboom (2023) framed MI in an even more causal language by stating that C should block all paths from any V to Y. That is, given the latent variable, all observed variables Y and covariates V are d-separated if MI holds.

In general, conditional independencies are testable implications in the data. The aforementioned sequential steps of MI testing have to be used because we cannot simply condition on the unobservable variable C. Its values can only be predicted (in the form of factor scores) by scores on the observed variables Y.

So far, we have kept our two parallel accounts of DAGs and the investigation of

MI rather abstract. To now show how (non-)invariance can be depicted by a DAG and to demonstrate how this can help to investigate MI in a more informative manner, we want to introduce an empirical example from moral psychology. In a multilab replication study, Bago et al. (2022) investigated which psychological and situational factors influence the judgement of moral dilemmas. They gathered data from 45 countries in all inhabited continents, leading to a final sample of $N = 22,112$ (after applying exclusion criteria like careless responding). For the following simulated and empirical demonstrations, we will use the *Oxford Utilitarianism Scale* (OUS; Kahane et al., 2018) from their paper. The OUS measures utilitarian thinking, that is, the notion that people's actions should always aim at maximizing the overall good. It comprises two independent subscales, *impartial beneficence* (IB; measured by 5 items) and *instrumental harm* (IH; measured by 4 items). IB describes the attitude that no individual is more important than another (e.g., "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal."), while IH entails that moral rules can be neglected if it is for a greater good (e.g., "It is morally right to harm an innocent person if harming them is a necessary means to helping several other innocent people"). To keep our examples illustrative, we only consider the measurement model of IB, which is a one-dimensional model with 5 items. The items are phrased as statements which are rated on a seven-point Likert scale (1 = "strongly disagree", 4 = "neither agree nor disagree", 7 = "strongly agree"). We refer interested readers to Kahane et al. (2018) for more details on the OUS.

To depict non-invariance by a DAG, we introduce another type of node, namely a *selection node* $\boxed{\text{S}}$. A selection node is not a variable but rather an indication for a group-specific distribution or causal relationship of the variable it is pointing into (Deffner et al., 2022; Pearl & Bareinboim, 2014). Thus, they are the key element when trying to incorporate non-invariance in a DAG. Assume that we want to test MI of the IB measurement model with respect to a binary covariate *Region*, defining group *western* and group *eastern*. We depict a group-specific distribution, that is, non-invariance of our observed variables Y by a selection node pointing into them, $\boxed{\text{S}} \rightarrow Y$ (see Figure 6).

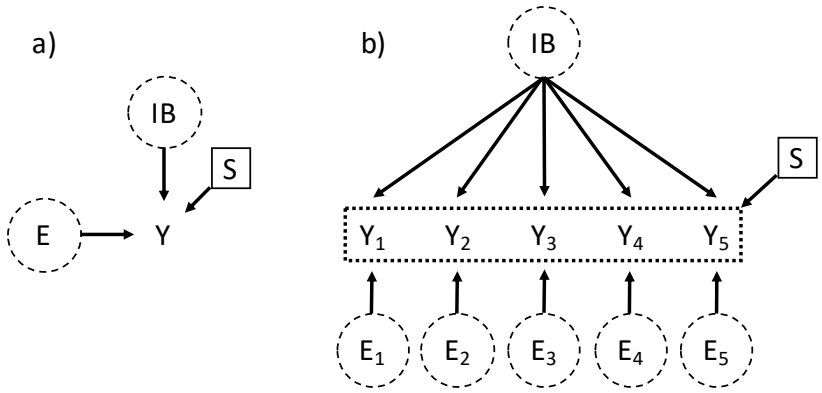

Figure 6
Study 3: DAG with a selection node pointing into the observed variables. a) Adaptation of Figure 6c in Deffner et al. (2022) where only one observed variable Y is shown; b) DAG of the complete measurement model of IB = impartial beneficence where the selection node points into potentially all observed variables Y_{1-5} (depicted by the dotted box around the observed variables).

Figure 6a) is similar to Figure 6c) in Deffner et al. (2022). However, they showed a latent variable with only one observed variable, which is not very common in psychological (questionnaire) assessment. In Figure 6b), the complete measurement model of *IB* is shown with a selection node pointing into potentially all observed variables Y_{1-5}. If one can make more detailed assumptions about group-specific selection mechanisms on the observed variables, the selection node could also only point into some, but not all, of the items. In the psychometric literature, this is often referred to as differential item functioning (Holland & Wainer, 2012; Zumbo, 2007). As Deffner and colleagues state, Figure 6 shows a selection node pointing into an outcome. This prevents unbiased comparisons of the observed (and consequently, the latent) variables between groups. Similar to what we mentioned in the introduction, an absent selection node is a stronger assumption than an existent one. Not drawing a selection node pointing into an observed variable encodes the assumption that this variable (here: questionnaire item) is invariant across *all* groups. In Figure 6, the selection node pointing into Y could subsume all four levels of non-invariance. By translating the DAG with a selection node (also called *selection diagram*) into a path diagram, we can see that one DAG implies many different models. In Figure 7, four different pairs (each consisting of group *western* and group *eastern*) of models are shown, where each pair depicts one level of MI being violated. The group-specific distribution of Y could stem from:

a) some paths between *IB* and Y being 0 in one group or a different number of latent variables between groups (configural non-invariance; Figure 7a),
b) the size of the loadings λ between *IB* and Y being different between groups (metric; Figure 7b),
c) the intercepts τ of Y being different between groups (scalar; Figure 7c),
d) or the variances of the unique errors E of Y being different between groups (residual; Figure 7d).

Now that we have introduced how to depict non-invariance with a selection diagram[15], we can turn to a more elaborate example. Specifically, we now demonstrate how DAGs can be used to make informed modeling decisions when investigating MI. We show how disregarding the complete causal model and instead only considering the groups that we want to compare, can miss important aspects of non-invariance. All code needed to reproduce the results of the following simulated and empirical example as well as a reproducible manuscript are available at https://osf.io/2mpq9/.

All analyses were conducted in the statistical software *R* (R Core Team, 2021), using the packages *lavaan* (Rosseel, 2012), *semTools* (Jorgensen et al., 2016), and *OpenMx* (Boker et al., 2011). The paper was written using the package *papaja* (Aust & Barth, 2020).

6.6 A More Holistic View on Measurement Invariance

We again consider our example in Figure 6b), that is, we want to compare the latent means of *IB* between groups western and eastern (defined by the covariate *Region*). To investigate whether scores of *IB* are comparable between these two groups, that is, if the measurement models are equivalent, we would first conduct a MG-CFA with *Region* as the grouping covariate. However, assume that the true data-generating process is not the one in Figure 6b) but the one in Figure 8, where an observed covariate *Age* is part of the measurement model.

In this setting, the selection node actually points into *Age*, not into the items Y_{1-5}. This means that not the distribution of Y_{1-5} varies between groups but the distribution

[15]The use of selection nodes to depict non-invariance highlights that —from a causal inference perspective— the concept of MI is related to *transportability*. We refer interested readers to Deffner et al. (2022) and Pearl and Bareinboim (2014) for more details.

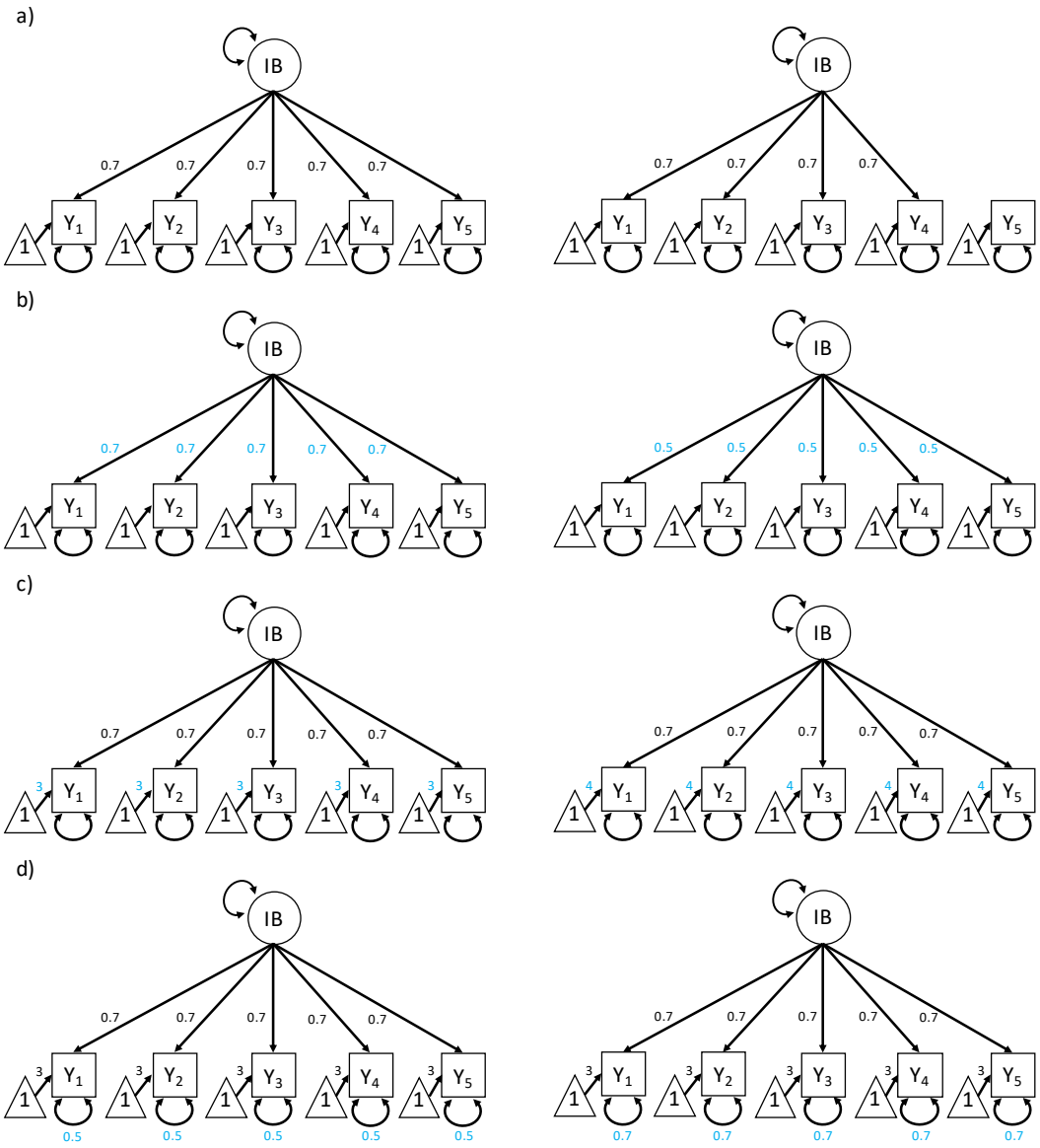

Figure 7

Study 3: Pairs of measurement models of IB (impartial beneficence) for which measurement invariance does not hold between the two groups. a) violation of configural invariance (violation of configural invariance due to different number of latent variables between groups is not displayed); b) violation of metric invariance (assuming standardized data); c) violation of scalar invariance; d) violation of residual invariance (assuming unstandardized data). Parameters that might differ between groups are highlighted in blue.

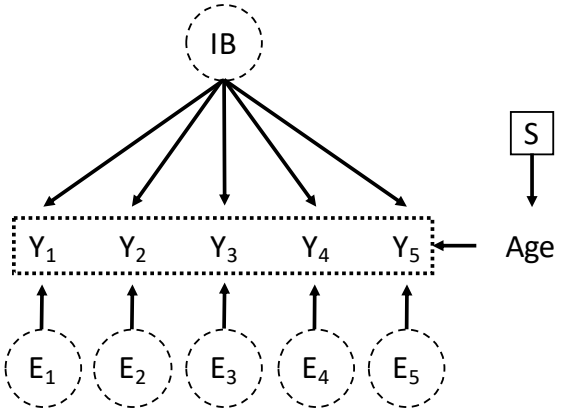

Figure 8
Study 3: DAG with a selection node pointing into the observed covariate Age which influences all observed variables Y_{1-5} (depicted by the dotted box around the observed variables).

of *Age*. Specifically, in group 1, $Age \sim N(0, 1)$ and in group 2, $Age \sim N(0.5, 1)$ (standardized ages where the mean age is higher in group 2 than in group 1). In this case, assume $IB \rightarrow Y \leftarrow Age$ to be an interaction between *IB* and *Age*, such that the measurement model for every item is $Y = (\sqrt{0.6} + 0.3Age) \cdot IB + E$ (cf. equation (15)).[16] That is, with increasing *Age*, the causal relationship between the latent variable *IB* and the observed variables Y_{1-5} grows stronger.

A small simulation of the model depicted in Figure 8[17] reveals the following: If we do not consider the DAG in Figure 8 and test MI following the current practice, that is, only test the invariance of measurement models between groups western and eastern, we find a significant violation of metric MI ($\chi^2(14; N = 1,000) = 19.46$, $p = .003$ and an increase in RMSEA of .028 for the comparison of the configural and the metric model). However, this result is only half of the picture: It is the different distribution of *Age* between groups that is decisive for the result of the MI test. That is, the group-specific mechanism, indicated by the selection node, is working on *Age*, not on the observed variables directly. Conclusions regarding different interpretations of the construct between groups western and eastern based on the MI test results are rather uninformative.

[16]Because DAGs do not impose a functional form on the relationships between variables, all variables jointly causing another variable can also interact (Deffner et al., 2022; Elwert, 2013).

[17]With $N = 1000$ ($n = 500$ per group), $IB \sim N(0, 1)$, and $diag(\Psi) \sim Uniform(0.2, 0.6)$. Together with loadings of $\sqrt{0.6}$, this results in an average item variance of 1.

How could DAGs have helped us to achieve more informative results regarding MI? Had we set up the selection diagram (by theoretical or empirical considerations) as in Figure 8, we would have seen that MG-CFA with *Region* as a grouping covariate is not the right model. Instead, we have to resort to a more flexible model to investigate MI in this case. We can read from the DAG that *Age* is an assumed direct cause of Y_{1-5} and that we assume *Age* to have a group-specific distribution. Thus, we want to include *Age* in our model in order to close the path between the selection node and the observed variables Y_{1-5} (remember that including *Age* in the model closes the path $\boxed{\text{S}} \rightarrow Age \rightarrow Y$). Generally speaking, our goal is to make as many assumptions as possible about covariates between the outcome (in our case Y) and the selection node, and then include these covariates in the model. This lets us gain more detailed information about group-specific mechanisms (i.e., non-invariance) in the data-generating process and how these mechanisms influence our observed variables.

One option to model the data-generating process depicted in Figure 8 is a type of moderated SEM called *moderated non-linear factor analysis* (MNLFA) (Bauer, 2017; Bauer & Hussong, 2009). MNLFA is especially suitable in this case because it allows the model parameters to depend on any covariate V in the data. In our example, we can model the expected loadings and intercepts by the regression equations $\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_0 + \boldsymbol{D}_{Region} Region_i + \boldsymbol{D}_{Age} Age_i$ and $\boldsymbol{\tau}_i = \boldsymbol{\tau}_0 + \boldsymbol{b}_{Region} Region_i + \boldsymbol{b}_{Age} Age_i$, respectively.[18] $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\tau}_0$ are the baseline loadings and intercepts, $\boldsymbol{D}$ and $\boldsymbol{b}$ are vectors[19] of linear effects of the covariates *Region* and *Age* on the parameters, and i denotes the person index. This model formulation allows us to estimate the baseline parameters as well as the individual effects a covariate has on the item parameters. Of course, if more detailed assumptions about which items are influenced by the covariates can be made, the equations above can be adjusted by setting the effects of the covariates on some items to 0. The covariate *Region* is also included to test its direct effect on the parameters (besides the assumed direct causal effect of *Age*). MNLFA can be estimated in *R* (R Core Team, 2021) via the package *OpenMx* (Boker et al., 2011). We refer readers to Kolbe et al. (2022) for a detailed guide on how to estimate MNLFA in

[18] Similarly, all other model parameters —like factor means or residual covariances— can be modeled as functions of covariates. Thus, MNLFA could be seen as a flexible extension to *multiple indicator multiple cause models* (*MIMIC models*; Muthén, 1989).

[19] In case of more than one latent variable, $\boldsymbol{D}$ would be a matrix with the same dimensions as $\boldsymbol{\Lambda}_0$.

OpenMx and specifically how to use it to investigate MI.

From a causal inference perspective, we can justify the model choice like this: the less we know about our measurement model and the covariates surrounding it, the more potential differences in parameters we have to consider during estimation and testing. The more potential differences we have to consider, the more arrows we should draw in our DAG.

6.6.1 Simulated Example

Table 27 shows the estimated results of a MNLFA for the simulated example described above. The model parameters (in our example: loadings and intercepts) are allowed to be moderated by covariates *Region* and *Age* as described above. This is the configural model. The advantage of modeling the assumed causal relationships like this is that we get detailed estimates of parameters and possible interactions for every item. As can be seen, *Region* does not have an influence on neither intercepts nor loadings, whereas *Age* has an influence of around 0.3 on the baseline loadings $\mathbf{\Lambda}_0$, which are around $\sqrt{0.6}$.

Table 27
Study 3: Results of moderated non-linear factor analysis for the toy example.

Item	τ_0	b_{Region}	b_{Age}	Λ_0	D_{Region}	D_{Age}
Item 1	-0.07	0.15	-0.05	1.04	-0.05	0.00
Item 2	-0.08	0.03	-0.03	0.80	-0.07	0.35
Item 3	-0.06	0.12	-0.04	0.75	0.01	0.27
Item 4	-0.09	0.06	-0.04	0.74	-0.03	0.30
Item 5	-0.04	0.04	-0.05	0.73	0.04	0.30

Note. τ_0 = Baseline intercepts, b_{Region} = (Additive) Effects of covariate Region on baseline intercepts, b_{Age} = (Linear) Effects of covariate Age on baseline intercepts, Λ_0 = Baseline loadings, D_{Region} = (Additive) Effects of covariate Region on baseline loadings, D_{Age} = (Linear) Effects of covariate Age on baseline loadings. Effects of Region and Age on other model parameters, e.g., residual variances, are not reported here. Reference category of Region is Eastern. The loading of item 1 was simulated as 1 for identification purposes.

Beyond visual inspection of the parameter estimates, we can also investigate metric and scalar MI. This is done by setting the effects of the covariates on the loadings (for metric MI), and loadings as well as intercepts (for scalar MI) to 0 and comparing these nested models. The results of this model comparison are shown in Table 28. They

show that metric MI is violated (by the covariate *Age*), whereas scalar MI is supported (i.e., there is no significant moderation of the intercepts by the covariates).

Table 28
Study 3: Results of χ^2-difference tests between the configural, metric, and scalar moderated non-linear factor analyses for the simulated example.

Comparison	$\Delta - 2LL$	Δdf	p-value
configural vs. metric	290.39	10.00	0.00
metric vs. scalar	-113.63	6.00	1.00

Note. $\Delta - 2LL$ = difference in -2 times the log-likelihood of the models, Δdf = Difference in degrees of freedom. A p-value of 0 means that it is < 0.005

By taking into account the whole causal model and using a more flexible method than simply relying on MG-CFA, we can make a more informed decision regarding MI. Had we only used MG-CFA, we would try to explain why the two regions western and eastern have non-invariant measurement models, which would be the wrong question. On the basis of theoretical and empirical assumptions regarding the causal relationships, however, we can now reason about why the relationship between the latent variable *IB* and its items grows stronger with increasing age. It should be highlighted again that drawing a DAG with many arrows and using MNLFA entails less assumptions (or assumptions that are less strong) than using MG-CFA with one covariate. From a causal inference perspective, MG-CFA could be seen as the MI testing approach with the most assumptions.

6.6.2 Empirical Example

To mimic the analysis of the simulated example in the example on the real data published by Bago et al. (2022), we only considered observations from group western whose age was above 30 years. This was done to achieve two approximately equally sized groups ($n_{western} = 2,911$; $n_{eastern} = 2,941$) with differing mean ages ($M_{western} = 43.22$; $M_{eastern} = 26.13$). Note that this changes the real data, which was done simply for didactic purposes; the following results should not be interpreted from a substantive research perspective.

Table 29 shows the results of a MG-CFA, where again a one-dimensional model is

specified and MI is investigated between the two groups western and eastern. We see that the results of the χ^2-difference test is statistically significant for the evaluation of both metric and scalar MI. This is an indication that neither of these two levels of MI hold, that is, neither loadings nor intercepts are equivalent across groups. Considering the RMSEA, the difference between the configural and metric model does not exceed commonly suggested cut-offs, therefore supporting metric MI (Chen, 2007; Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014). The RMSEA difference between the metric and the scalar model again indicates a violation of scalar MI. Based on these results, all we can conclude for now is that MI does not hold between the two regions western and eastern.

Table 29
Study 3: Results of multi-group confirmatory factor analysis for the empirical example between regions western and eastern.

Model	df	χ^2	$\Delta\chi^2$	Δdf	p-value	RMSEA
configural	10	104.36	-	-	-	0.06
metric	14	142.72	38.36	4	0.00	0.06
scalar	18	373.54	230.83	4	0.00	0.08

Note. df = Degrees of freedom, χ^2 = Value of the test statistic, $\Delta\chi^2$ = Difference in values of the test statistics, Δdf = Difference in degrees of freedom, RMSEA = Root mean square error of approximation. A p-value of 0 means that it is < 0.005.

To be able to reason more about the role of non-invariance in the underlying data-generating process, we again have to consider the complete DAG and model the data-generating process accordingly. Table 30 shows the results of a MNLFA, where both covariates *Age* and *Region* are allowed to moderate the parameter estimates. In the empirical example, these results paint a different picture than before. *Age* has no effect on both loadings and intercepts, whereas *Region* directly influences (primarily) the item intercepts. Specifically, in the group western, the intercepts of items 1, 2, and 4 are higher compared to group eastern, whereas for item 3, the intercept is lower. Effects of *Region* on the loadings are less strong. Similar to the simulated example, a χ^2-difference test can be conducted. By this we can test whether allowing that the parameters are moderated by the covariates *Region* and *Age* significantly increases model fit (and thus, whether MI is violated).

Table 30
Study 3: Results of moderated non-linear factor analysis for the empirical example.

Item	τ_0	b_{Region}	b_{Age}	Λ_0	D_{Region}	D_{Age}
Item 1	3.68	0.22	0.00	0.96	-0.16	0.00
Item 2	3.21	0.24	0.01	1.41	0.02	-0.01
Item 3	4.26	-0.19	0.01	0.68	0.12	0.00
Item 4	2.77	0.49	0.02	0.82	0.10	0.00
Item 5	3.42	0.05	0.01	1.17	-0.10	0.00

Note. τ_0 = Baseline intercepts, b_{Region} = (Additive) Effects of covariate Region on baseline intercepts, b_{Age} = (Linear) Effects of covariate Age on baseline intercepts, Λ_0 = Baseline loadings, D_{Region} = (Additive) Effects of covariate Region on baseline loadings, D_{Age} = (Linear) Effects of covariate Age on baseline loadings. Effects of Region and Age on other model parameters, e.g., residual variances, are not reported here. Reference category of Region is Eastern.

Table 31 shows that both levels of MI, metric and scalar, are violated. That is, the covarirates *Region* and *Age* significantly influence the loadings and intercepts in our measurement model. Because the model outputs estimates for all item parameters and their moderators, we are able to reason in more detail about the causes of non-invariance, given our assumptions encoded in the DAG. Of course, detailed inspection of item contents would now be necessary to explain why a covariate influences the item parameters. Since this would be beyond the scope of this paper and since we are not subject matter experts in moral psychology, we end our empirical demonstration here. However, we hope that this example proves as a starting point for showing how MI can be investigated according to the underlying causal assumptions.

Table 31
Study 3: Results of χ^2-difference tests between the configural, metric, and scalar moderated non-linear factor analyses for the empirical example.

Comparison	$\Delta - 2LL$	Δdf	p-value
configural vs. metric	44.66	10.00	0.00
metric vs. scalar	276.26	6.00	0.00

Note. $\Delta - 2LL$ = difference in -2 times the log-likelihood of the models, Δdf = Difference in degrees of freedom. A p-value of 0 means that it is < 0.005

6.7 Discussion

In this paper, we first introduced the connection between DAGs used in causal inference and path diagrams of measurement models, which are more common in the psychometric literature. We then showed how a lack of MI can be depicted by a DAG. We demonstrated how taking into account the causal relationships between the measurement model and the surrounding covariates yields more informative results when investigating MI. If MI is directly violated by a covariate that is not of primary interest (e.g., age in our example above), DAGs can help to visualize the underlying assumptions. Specifically, they depict the assumed mechanisms by which the data-generating process differs between groups. In this, researchers can find appropriate statistical models like MNLFA that allow them to estimate an extended measurement model. This also lets us reason about the causes of non-invariance. Only by investigating *why* MI does not hold, we can see it as an important finding by itself and draw conclusions about how different groups interpret a construct (Putnick & Bornstein, 2016).

One critique against DAGs is that it is difficult to specify all causal relationships (surrounding the measurement model, in our case). This is true but we deem this an argument against poor psychological theories and not against DAGs. A sound theory should allow us to specify the relationships between the variables it comprises. Besides, as mentioned in the introduction, also an incomplete or even wrong DAG can help us to reveal specific issues in theories. For example, drawing a DAG and realizing that there is uncertainty regarding some relationships, can be the starting point of further scientific discourse. In the end, DAGs are not about adding assumptions — they are about revealing the assumptions that are otherwise made implicitly (Deffner et al., 2022; Pearl & Bareinboim, 2014).

DAGs and path diagrams are part of a broader class of graphical models that have been introduced in the psychometric literature. Other examples are graphical Rasch models and graphical regression models that explicitly depict and model differential item functioning or local dependence (i.e., correlated item responses even after conditioning on the latent variable) (Anderson & Böckenholt, 2000; Kreiner & Christensen, 2002, 2011). Similarly, latent class models have been visualized as categorical causal models, again facilitating the representation of underlying model assumptions, such as local

independence (Bartolucci & Forcina, 2005; Hagenaars, 1998; Humphreys & Titterington, 2003; Rijmen et al., 2008). In this notion, local dependence is intertwined with unobserved confounding (i.e., failing to include a covariate that influences the item response in the measurement model).

6.7.1 Limitations and Future Research

Our goal was to provide a translation between path diagrams of measurement models and DAGs, thereby framing MI and its investigation as a causal inference problem. In this, we showed only one example with one observed covariate (i.e., age with different distributions between groups). Needless to say, many more causal relationships leading to a violation of MI are conceivable, for example one in which the cause of non-invariance is latent. A prominent example of this in the literature is acquiescence bias, that is, the tendency of respondents to agree more to statements or items, irrespective of the content of the item (D'Urso et al., 2023; Lechner et al., 2019). Even further, beyond the representation of latent variables as common causes of observed variables, DAGs might help to depict (non-)invariance in other representations of multivariate data. Most notably, network models have been proposed as such an alternative conceptualization (Borsboom et al., 2021), and this field is increasingly interested in the investigation of invariance of networks across groups (e.g., Hoekstra et al., 2023). In these cases, graphical tools from the causal inference literature might also aid to reason about the causes of non-invariance and to find appropriate approaches with which the causal relationships can be modeled. Future studies could therefore illustrate the usefulness of DAGs when investigating MI in different scenarios or conceptualizations.

6.8 Conclusion

Many psychological studies concern some comparison of latent scores between groups. Investigating whether measurement models of the latent variables are equivalent between groups is crucial for unbiased conclusions. We discussed a theoretical framework in which MI can be viewed from a causal inference perspective. Reasoning about causes of differences in how constructs and their measures function across groups can create valuable insights for scale construction or even theory building. Drawing a DAG which encodes assumptions about non-invariance helps researchers to make informed modeling

choices. In this, it might encourage them to view MI as part of the modeling process and as an interesting topic of research by itself — and not just as an additional test prior to the actual data analysis. Ultimately, we hope to contribute to an increase in the prevalence of investigations of MI.

6.9 References

Anderson, C. J., & Böckenholt, U. (2000). Graphical regression models for polytomous variables. *Psychometrika*, *65*(4), 497–509. https://doi.org/10.1007/BF02296340

Asparouhov, T., & Muthén, B. O. (2014). Multiple-Group Factor Analysis Alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495–508. https://doi.org/10.1080/10705511.2014.919210

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

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

Bartolucci, F., & Forcina, A. (2005). Likelihood inference on the underlying structure of IRT models. *Psychometrika*, *70*(1), 31–43. https://doi.org/10.1007/s11336-001-0934-z

Bastian, B., Kuppens, P., De Roover, K., & Diener, E. (2014). Is valuing positive emotion associated with life satisfaction? *Emotion*, *14*, 639–645. https://doi.org/10.1037/a0036466

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

Bauer, D. J., Belzak, W. C. M., & Cole, V. T. (2020). Simplifying the Assessment of Measurement Invariance over Multiple Background Variables: Using Regularized Moderated Nonlinear Factor Analysis to Detect Differential Item Functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 43–55. https://doi.org/10.1080/10705511.2019.1642754

Bauer, D. J., & Hussong, A. M. (2009). Psychometric approaches for developing commensurate measures across independent studies: Traditional and new models. *Psychological Methods*, *14*(2), 101–125. https://doi.org/10.1037/a0015583

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods*, *25*(6), 673–690. https://doi.org/10.1037/met0000253

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., Spies, J., Estabrook, R., Kenny, S., Bates, T., Mehta, P., & Fox, J. (2011). OpenMx: An Open Source Extended Structural Equation Modeling Framework. *Psychometrika*, *76*(2), 306–317. https://doi.org/10.1007/s11336-010-9200-6

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Bollen, K. A., & Pearl, J. (2013). Eight Myths About Causality and Structural Equation Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 301–328). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_15

Borsboom, D. (2023). Psychological Constructs as Organizing Principles. In L. A. van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on Contemporary Psychometrics* (pp. 89–108). Springer International Publishing. https://doi.org/10.1007/978-3-031-10370-4_5

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., Robinaugh, D. J., Perugini, M., Dalege, J., Costantini, G., Isvoranu, A.-M., Wysocki, A. C., Borkulo, C. D. van, Van Bork, R., & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*(1), 1–18. https://doi.org/10.1038/s43586-021-00055-w

Brandmaier, A. M., Oertzen, T. von, McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement

invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

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

D'Urso, E. D., Tijmstra, J., Vermunt, J. K., & De Roover, K. (2023). Does Acquiescence Disagree with Measurement Invariance Testing? *Structural Equation Modeling: A Multidisciplinary Journal*, *0*(0), 1–15. https://doi.org/10.1080/10705511.2023.2260106

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

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A Causal Framework for Cross-Cultural Generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3). https://doi.org/10.1177/25152459221106366

Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Springer Netherlands. https://doi.org/10.1007/978-94-007-6094-3_13

Epskamp, S. (2015). semPlot: Unified Visualizations of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 474–483. https://doi.org/10.1080/10705511.2014.937847

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, Online First. https://doi.org/10.1177/00131644231163813

Greenland, S., & Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, *31*(5), 1030–1037. https://doi.org/10.1093/ije/31.5.1030

Hagenaars, J. A. (1998). Categorical Causal Modeling: Latent Class Analysis and Directed Log-Linear Models with Latent Variables. *Sociological Methods &*

Research, *26*(4), 436–486. https://doi.org/10.1177/0049124198026004002

Hoekstra, R. H. A., Epskamp, S., Nierenberg, A., Borsboom, D., & McNally, R. J. (2023). *Testing similarity in longitudinal networks: The Individual Network Invariance Test (INIT)*. https://doi.org/10.31234/osf.io/ugs2r

Holland, P. W., & Wainer, H. (2012). *Differential item functioning*. Routledge.

House, B. R., Kanngiesser, P., Barrett, H. C., Broesch, T., Cebioglu, S., Crittenden, A. N., Erut, A., Lew-Levy, S., Sebastian-Enesco, C., Smith, A. M., Yilmaz, S., & Silk, J. B. (2020). Universal norm psychology leads to societal diversity in prosocial behaviour and development. *Nature Human Behaviour*, *4*(1), 36–44. https://doi.org/10.1038/s41562-019-0734-z

Humphreys, K., & Titterington, D. M. (2003). Variational approximations for categorical causal modeling with latent variables. *Psychometrika*, *68*(3), 391–412. https://doi.org/10.1007/BF02294734

Jöreskog, K. G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika*, *32*(4), 443–482.

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

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., Garnier-Villarreal, M., Selig, J., Boulton, A., Preacher, K., et al. (2016). Package "semtools." *Https://Cran.r-Project.org/Web/Packages/semTools/semTools.pdf*.

Kahane, G., Everett, J. A. C., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*(2), 131–164. https://doi.org/10.1037/rev0000093

Kim, E., Cao, C., Wang, Y., & Nguyen, D. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling A Multidisciplinary Journal*, *24:4*, 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2022). Assessing measurement invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, *Advance Online Publication*. https://doi.org/10.1037/met0000501

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor Selection Strategies for DIF Analysis: Review, Assessment, and New Approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. https://doi.org/10.1177/0013164414529792

Kreiner, S., & Christensen, K. B. (2002). Graphical Rasch Models. In M. Mesbah, B. F. Cole, & M.-L. T. Lee (Eds.), *Statistical Methods for Quality of Life Studies: Design, Measurements and Analysis* (pp. 187–203). Springer US. https://doi.org/10.1007/978-1-4757-3625-0_15

Kreiner, S., & Christensen, K. B. (2011). Item Screening in Graphical Loglinear Rasch Models. *Psychometrika*, *76*(2), 228–256. https://doi.org/10.1007/s11336-011-9203-y

Kunicki, Z. J., Smith, M. L., & Murray, E. J. (2023). A Primer on Structural Equation Model Diagrams and Directed Acyclic Graphs: When and How to Use Each in Psychological and Epidemiological Research. *Advances in Methods and Practices in Psychological Science*, *6*(2). https://doi.org/10.1177/25152459231156085

Lechner, C. M., Partsch, M. V., Danner, D., & Rammstedt, B. (2019). Individual, situational, and cultural correlates of acquiescent responding: Towards a unified conceptual framework. *British Journal of Mathematical and Statistical Psychology*, *72*(3), 426–446. https://doi.org/10.1111/bmsp.12164

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

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. IAP.

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

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

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5

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

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L., Billiet, J., & Schmidt, P. (2022). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research*, *52*(3), 1401–1419. https://doi.org/10.1177/00491241221091755

Mulaik, S. A. (2009). *Linear Causal Modeling with Structural Equations*. CRC Press.

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

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods & Research*, *27*(2), 226–284. https://doi.org/10.1177/0049124198027002004

Pearl, J. (2012). *The Causal Foundations of Structural Equation Modeling:* Defense Technical Information Center. https://doi.org/10.21236/ADA557445

Pearl, J., & Bareinboim, E. (2014). External Validity: From Do-Calculus to Trans-

portability Across Populations. *Statistical Science*, *29*(4). https://doi.org/10.1214/14-STS486

Pohl, S., Schulze, D., & Stets, E. (2021). Partial Measurement Invariance: Extending and Evaluating the Cluster Approach for Identifying Anchor Items. *Applied Psychological Measurement*, *45*(7-8), 477–493. https://doi.org/10.1177/01466216211042809

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

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent Class Models for Diary Method Data: Parameter Estimation by Local Computations. *Psychometrika*, *73*(2), 167–182. https://doi.org/10.1007/s11336-007-9001-8

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

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

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*, 31–57. https://doi.org/10.1177/0013164413498257

Sass, D. A. (2011). Testing Measurement Invariance and Comparing Latent Factor Means Within a Confirmatory Factor Analysis Framework. *Journal of Psychoeducational Assessment*, *29*(4), 347–363. https:

//doi.org/10.1177/0734282911406661

Schulze, D., & Pohl, S. (2021). Finding Clusters of Measurement Invariant Items for Continuous Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 219–228. https://doi.org/10.1080/10705511.2020.1771186

Seifert, I. S., Rohrer, J. M., & Schmukle, S. C. (2024). Using within-person change in three large panel studies to estimate personality age trajectories. *Journal of Personality and Social Psychology*, *126*(1), 150–174. https://doi.org/10.1037/pspp0000482

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

Sterner, P., & Goretzko, D. (2023). Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30:6*, 871–886. https://doi.org/10.1080/10705511.2023.2188573

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch Trees: A New Method for Detecting Differential Item Functioning in the Rasch Model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Suzuki, E., Shinozaki, T., & Yamamoto, E. (2020). Causal Diagrams: Pitfalls and Tips. *Journal of Epidemiology*, *30*(4), 153–162. https://doi.org/10.2188/jea.JE20190192

Tutz, G., & Schauberger, G. (2015). A Penalty Approach to Differential Item Functioning in Rasch Models. *Psychometrika*, *80*(1), 21–43. https://doi.org/10.1007/s11336-013-9377-6

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

Van De Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492.

https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Westfall, J., & Yarkoni, T. (2016). Statistically Controlling for Confounding Constructs Is Harder than You Think. *PLOS ONE*, *11*(3), e0152719. https://doi.org/10.1371/journal.pone.0152719

Wysocki, A. C., Lawson, K. M., & Rhemtulla, M. (2022). Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*, *5*(2). https://doi.org/10.1177/25152459221095823

Zumbo, B. D. (2007). Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/10.1080/15434300701375832

7 General Discussion

The three manuscripts in this thesis addressed various methodological topics on MI. Modern statistical approaches were employed to introduce new methods and concepts that improve investigations of MI. Study 1 presented EFA trees, a new EFA-based method to assess MI in the earliest stages of questionnaire development. Study 2 provided an overview and comparison of EFA-based methods. By means of an empirical example in three different statistical software programs, Study 2 hopefully facilitates the application of these EFA-based methods. Study 3 proposed a framework based on causal inference that allows researchers to visualize and reason about potential causes of non-invariance. By using the framework to depict causal structures underlying the investigation of MI, non-invariance can be viewed as a substantively interesting topic of research itself.

A detailed discussion of the results and limitations of the three studies can be found in the respective chapters. In the following, I want to focus on topics beyond this thesis. First, I critically discuss limitations regarding the general applicability of the methods and the framework presented in this thesis. Second, I provide an outlook on topics for future research on MI; specifically, tailored fit index cut-offs, effect size measures, and longitudinal MI. Third and last, I address recent criticism against the necessity of MI as a prerequisite for latent mean comparisons.

7.1 Solutions in Search of a Problem?

In the following, I want to critically discuss the utility and applicability of the proposed methods and framework for applied psychological science — arguably the goal of methodological research. Throughout this thesis, I have mentioned repeatedly that MI is rarely investigated in practice although methodological research has shown its importance when comparing, for example, latent means (Maassen et al., 2023). Thus, there is a discrepancy between what methodologists recommend and what applied researchers do when it comes to investigating MI. I have already hinted at two possible reasons for this mismatch: the difficulty for applied researchers to fully grasp the vast variety of methods to investigate MI and the notion that MI is merely a statistical assumption that one has to test to license further analyses — rather than a substantively

interesting topic by itself. In that sense, the manuscripts of this thesis seem useful for applied researchers because they address these issues.

However, our manuscripts are not the first to provide an overview and comparison of methods (Kim et al., 2017; Luong & Flake, 2023) or aim at embedding statistical concepts in a more applied framework (Van Bork et al., 2022). Thus, as they currently stand in the literature, the present manuscripts require further efforts to increase the prevalence of investigations of MI in the psychological literature. Study 1 is the first paper to introduce EFA trees and complex statistical procedures but without extensive empirical examples and provides only template code as software implementation of the method. Study 2 is the first paper to contrast newly developed EFA-based methods and gives some of these methods their first ever empirical application. Yet still, this application only demonstrates the methods at hand and, thus, lacks direct connection to substantive theory. It does, however, improve Study 1 by introducing an R package for an easier implementation of EFA trees, thus providing a remedy for the software implementation problem. Study 3 paints a broader picture by shifting the focus from a purely statistical investigation of MI to a causal assessment of possible reasons of non-invariance. While this may motivate researchers to approach the issue of (non-)invariance from a theoretically informed perspective, an end-to-end example based on substantive theory of how this could be done is needed to increase its applicability for applied researchers.

To summarize, the manuscripts of this thesis provide important steps forward to tackle issues of investigating MI. However, as I outline below, further work is needed to make them readily applicable for applied researchers from various psychological disciplines. This is especially important to reduce the discrepancy between what methodologists recommend and what applied researchers do when it comes to investigating MI. The mismatch between methodological developments and empirical applications of statistical methods is neither new nor unique to the topic of MI. Borsboom (2006a) extensively discussed the discrepancy between psychometrics and psychology, criticizing that psychometric developments are not integrated into psychological research. Sijtsma (2006) acknowledged this problem but takes a more lenient position by stating that psychometric models (as a result of psychometric research) are simply a tool for data

analysis (see also Sijtsma, 2012). Consequently, he concluded that applied researchers are most likely to adopt newly developed methods if they are convinced of their superiority over classical methods. Additionally, new developments should be easy to implement.

I fully agree with Sijtsma (2006). Methodological development does not stop once a method is "developed"; that is, when it is published in a methodological journal. Rather, it should be seen as a process. I see at least three important steps that follow the step of introducing a new method: dissemination through tutorials, software implementation, and collaboration. First, once a method has been introduced, it should be the methodologist's goal to disseminate statistical findings to applied researchers. By writing tutorial papers, methodologists could demonstrate the assumptions needed to apply a method and the conclusions drawn from its output. This should be done repeatedly by publishing widely in different fields, as different fields have different challenges in analyzing data or applying models (Borsboom, 2006a). There are excellent examples of tutorial and review papers that make methodological concepts, like MI, comprehensible to applied researchers. These are either tailored to a specific field of psychology (Putnick & Bornstein, 2016; Somaraju et al., 2022; Vandenberg, 2002; Vandenberg & Lance, 2000) or to a specific method (Kolbe et al., 2022; Luong & Flake, 2023). Psychological science can only benefit from such work. To advance in the process of methodological development of the methods and framework presented in this thesis, tutorial papers are needed to disseminate their application and usefulness. Second, the likelihood of newly developed methods being applied increases when they are available in (ideally open-source) statistical software (Borsboom, 2006a; Sijtsma, 2006). Thus, similar to the theoretical dissemination through tutorial papers, the implementation of newly developed methods in statistical software is an important step to increase practical dissemination. Ideally, these two steps are combined, by including software code in the tutorial papers (see e.g., Pargent et al., 2023). Third, and perhaps most importantly, method development and application should not be seen as two different pairs of shoes. Instead, both parts can be integrated, either by psychometricians applying their models to solve substantive problems (Borsboom, 2006a) or by psychometricians and applied researchers collaborating (Sijtsma, 2006). This allows methodological research to identify actual issues in applied research that

need to be addressed.

Much of psychology is a quantitative science. The majority of studies employ psychometric models to empirical data to find statistical evidence for verbal claims. Methodological development is therefore crucial to bring psychology forward as a science. However, it is equally important for methodologists, in general, and psychometricians, in particular, to maintain a connection to applied research. Especially for investigations of MI, and thus also for the methods and framework of this thesis, it is important to be continuously disseminated and implemented in the applied literature to close the gap between methodological recommendations and applied reality (Maassen et al., 2023) — be it through tutorials, software implementations, or collaborations. This ensures that methodological research is not developing statistical solutions in search of an empirical problem.

7.2 Future Research

7.2.1 Tailored Fit Index Cut-Offs

As discussed in the Introduction of this thesis, MI is primarily investigated by means of χ^2-difference tests and changes in fit indices between increasingly restricted models (Putnick & Bornstein, 2016; Van de Schoot et al., 2012). When deciding whether a decrease or increase in fit indices is indicative of non-invariance, researchers usually draw on fixed cut-offs, for example an increase in RMSEA of 0.01 or a decrease in CFI of 0.01. Some more nuanced recommendations exist that take into account specific factors influencing the sensitivity of these indices to violations of MI; for example, in the context of many groups (Rutkowski & Svetina, 2014) or when models contain cross-loadings (Cao & Liang, 2022a). This is to be applauded because it allows for a more accurate assessment of MI, catered to a specific setting. Nonetheless, the fact that the same cut-offs do not hold for all research conditions and that more fine-grained recommendations are even necessary raises the question whether (recommendations of) fixed cut-offs should be applied in the first place. It has been shown repeatedly that appropriate cut-offs to quantify "good" or "adequate" model fit depend on many different factors; these include model size, sample size, various nuisance parameters like loading size, and different types of potential model misspecifications, to name just

a few (Goretzko et al., 2023; Heene et al., 2011, 2012; Partsch et al., 2024; Savalei, 2012; Savalei et al., 2023). This is because all of these factors influence the sensitivity of fit indices to identify model misspecifications, making it impossible to reasonably apply the same cut-off to different data sets and models. Since a lack of MI can be seen as a form of model misspecification in a MG-CFA model (e.g., equal loadings are not tenable across groups), these findings are also relevant for investigations of MI by means of changes in fit indices.

To address this issue in single-group settings, researchers developed ways to "tailor" cut-offs of fit indices to the model and data conditions at hand. Specifically, McNeish and Wolf (2023) developed the *dynamic fit index cut-offs*, Schmalbach et al. (2019) the *ezCutoffs*, and Groskurth et al. (2022) a method involving a receiver–operating characteristic (ROC) curve analysis (for an explanation and comparison of these methods, see Goretzko et al., 2023). All of these three methods are simulation-based, allowing to take into account the specific model and data characteristics. In this, more appropriate cut-offs for real-life applications can be derived. At the same time, they preserve the possibility to categorically assess the fit of a model, for example, classifying the fit as "good" or "adequate".

Especially the dynamic fit index cut-offs by McNeish and Wolf (2023) have received many extensions, making them broadly applicable. Most notably, they have been extended to provide tailored cut-offs for ordinal and binary item responses (McNeish, 2023) as well as for any covariance structure model, among others bifactor models (McNeish & Wolf, 2024). As mentioned by McNeish and Wolf (2023), an important project for future research would be to extend the dynamic fit index cut-offs to investigations of MI. That is, instead of relying on the commonly used and suggested cut-offs (e.g., a decrease in CFI of 0.01), cut-offs tailored to a specific application could be provided to indicate whether parameter restrictions (e.g., equal loadings across groups) worsen the fit of a model. As in single-group settings, researchers would have to specify a configural MG-CFA (hypothesized model) and use this model to create an alternative, slightly misspecified version (e.g., a MG-CFA with a cross-loading in only one group). M data sets are then simulated twice, once with the hypothesized and once with the alternative model as the data-generating model. The hypothesized model

could be fit to both sets of M data sets which yields two distributions of fit indices. The first one represents the distribution of fit indices under correct specification because the model is fit to the data that were generated using itself as the true model. The second one represents the distribution of fit indices under misspecification because the model is fit to the data stemming from a different (non-invariant) data-generating model. These two distributions could then be used to derive cut-offs with adequate false-positive and false-negative rates (for more details, see McNeish & Wolf, 2023). Depending on which model is used as the hypothesized model (configural, metric, scalar, or residual model), cut-offs for the respective level of MI could be derived. Such dynamic fit index cut-offs for MI would allow researchers to reach a binary decision regarding (non-)invariance. At the same time, they take into account that specific model and data characteristics influence the sensitivity of fit indices to detect non-invariance (Cao & Liang, 2022a, 2022b; Chen, 2007; Cheung & Rensvold, 2002). It should be kept in mind, however, that this new way of determining a threshold to diagnose non-invariance does not eleviate the issue of handling non-invariance. We still need appropriate ways of dealing with a lack of MI, for example by using the causal framework from Study 3.

7.2.2 Effect Size Measures

As just described, MI is currently primarily assessed by changes in fit indices (e.g., CFI or RMSEA) or by means of χ^2-difference tests (Putnick & Bornstein, 2016). Strictly speaking, these fit indices and even the test statistic of a χ^2-difference test can be seen as an effect size measure (McNeish & Wolf, 2023). However, with regard to meaningful conclusions about the impact of non-invariance on substantive analyses, both approaches have downsides. As just discussed, fit indices are heavily influenced by model and data conditions (Goretzko et al., 2023; Heene et al., 2011; Partsch et al., 2024). It is difficult to assess let alone quantify the influence of non-invariance on statistical inference just by changes in fit indices alone. An increase in RMSEA or a decrease in CFI cannot be directly related to a distortion of p-values or regression parameters in stastical inference. χ^2-difference tests assess the hypothesis of exact parameter equivalence, that is, a parameter difference between groups of exactly zero. However, in every analysis, the actual question we should ask is at which degree non-invariance distorts inference in our substantive analyses, like latent mean comparisons

(Borsboom, 2006b; Gunn et al., 2020). Put differently, we want to assess whether small, but non-zero, parameter differences would even matter for our analyses. This raises the need for effect size measures that provide information whether non-invariance is large enough to distort statistical inference, instead of just being significantly non-zero (Funder & Gardiner, 2024).[20]

To address this issue, Nye and Drasgow (2011) introduced the effect size d_{MACS} (MACS: Mean and Covariance Structure). d_{MACS} estimates the degree of non-invariance for each item in a standardized metric (in the sense of standard deviations). Thus, the magnitude of the effect of non-invariance can be quantified, independent of the sample size. Additional analyses can then reveal the practical consequences of non-invariance on parameters like the mean or variance of scores of a scale. One limitation of d_{MACS} is that it is only applicable to models with an independent clusters structure. Items with cross-loadings cannot be evaluated because these additional loadings are not included in the formula of d_{MACS} but would impact the predicted item response (Nye & Drasgow, 2011). This makes d_{MACS} less ideal when multi-dimensional questionnaires are investigated for MI with EFA-based methods, which specifically consider and estimate (differences in) cross-loadings. In these cases, it could only be applied to items for which cross-loadings are estimated to be (close to) zero. Gunn et al. (2020) have already extended d_{MACS} to make it more widely applicable, for example to categorical item responses, but did not consider cross-loadings either. Consequently, for effect size measures to join the trend of EFA-based assessments of MI, future research could generalize d_{MACS} to models with cross-loadings.

Another effect size that has been proposed is the *expected parameter change*-interest (EPC-interest; Oberski, 2014; Oberski et al., 2015). It quantifies the impact of freeing parameters that have been restricted on parameters of interest for the substantive analysis. For example, how does releasing the restriction of equal loading(s) across groups change the latent means? In this, EPC-interest can be seen as a sensitivity analysis that does not tell us which parameters are invariant but whether violations of MI change the conclusions of our substantive analyses (Oberski et al., 2015). By

[20] Of course, even with standardized effect sizes available, the question would remain what effect magnitude would be substantially meaningful. Nonetheless, this question could be addressed more adequately from a content related perspective than the question whether, for example, a difference in fit indices is meaningful (e.g., similar to interpretations of Cohen's δ).

investigating the impact of specific parameter restrictions on parameters of interest, the model can also be modified, similar to modification indices. But as Oberski (2014) notes, this strategy inherits the problem of capitalization on chance (and thus, lacks generalizability; MacCallum et al., 1992) and should not replace theoretical considerations when altering a model. EPC-interest was introduced considering only metric and scalar invariance, while assuming configural invariance. However, violations of configural invariance, like cross-loadings, can affect structural parameters (of interest), like latent means or regression coefficients (Oberski, 2014). Thus, similar to d_{MACS}, future research should investigate the applicability of EPC-interest to models with cross-loadings (i.e., in EFA-based investigations of MI).

7.2.3 Longitudinal Measurement Invariance

This thesis and the manuscripts it comprises focused on the invariance of measurement models between cross-sectional groups, for example different regions. However, MI also concerns the invariance of measurements across time, that is, across subsequent measurement occasions (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). For example in clinical psychology, it is common to compare pre- and post-therapy depression scores, sometimes even with additional follow-up measurements at a later time point (e.g., Fokkema et al., 2013). If we want to conclude that changes in depression scores occurred only due to true changes in the underlying construct, and not due to changes in measurement properties, we have to make sure that depression was assessed equivalently at each time point. This longitudinal view on MI poses an additional challenge regarding the conceptualization of MI. An interesting topic of future research would be to tackle this challenge using the causal framework introduced in Study 3: how can we assess longitudinal MI if changes in the interpretation of a construct between measurements are expected or even desired? In these cases, we would have to disentangle the different causal influences on item responses over time, for example, actual changes in the construct, possible interventions between measurements, and (potentially time-variant) covariates, like gender or age.

Consider a possible depression item targeting somatic complaints, for example some form of physical pain (e.g., Rush et al., 1996). Somatic complaints can be, but do not necessarily have to be, a symptom of depression according to the 11^{th} *revision of the Inter-*

national Classification of Diseases (World Health Organization, 2022). Thus, a positive relation between (latent) depression severity and the item response pre-treatment may be plausible (i.e., a positive loading). If we now treat a person with psychotherapy (targeting the depression severity) and supplement the psychotherapy with medication (specifically targeting the somatic complaints), the item response might change when being measured again post-treatment. This change, however, is not necessarily only related to the psychotherapy reducing depression severity. That is, the causal path to somatic complaints is not only $Psychotherapy \rightarrow Depression\ Severity \rightarrow Somatic\ Complaints$ but also $Medication \rightarrow Somatic\ Complaints$. As a result, the criterion of MI is violated: some covariates (here: medication) are no longer independent of the item responses (here: somatic complaints rating) given the latent variable (here: depression severity). This in turn can change the factorial structure of a questionnaire across time points, for example when the dose of medication changes the loading of the item asking about somatic complaints. Other items that assess actual symptoms of depression, like loss of pleasure, might also be answered differently but without changing the factorial structure. A potential change in item response would simply (and ideally only) be caused by changes in latent depression severity.[21]

In fact, Fried et al. (2016a) found that for four commonly used self- and clinician report depression measures, the factorial structure changed over time. In other words, longitudinal MI did not hold. While Fried et al. (2016a) extensively discussed possible reasons for this response shift (e.g., decrease of variability of symptoms across time), they were not able to "identify the culprit" (p. 1364). Similarly, Fokkema et al. (2013) found that post-treatment assessment of depression showed different measurement properties compared to pre-treatment assessments. Specifically, the post-treatment item scores seemed to overestimate depressive symptoms (scalar non-invariance) and measurement errors were smaller (residual non-invariance). This might indicate that patients got better at understanding and evaluating their own symptoms (Fokkema et al., 2013), which is to be expected and even desired because psychoeducation is a

[21] One could also question the underlying conceptualization of a reflective measurement model. In this case, the measurement model would have to be changed, for example to a network model of interrelated symptoms (Borsboom et al., 2021; Fried et al., 2016b; 2020). Nonetheless, if relations in a network change over time, the question remains what caused these changes and how we can make sense of differences in networks (or any other models) across different time points, potentially including covariates that influence these changes.

common element of psychotherapy (Donker et al., 2009; Tursi et al., 2013).

I am convinced that incorporating these phenomena into the causal framework of Study 3 can help us to better understand how constructs and measures thereof function over time. By depicting our assumptions in a DAG, we can make informed modeling decisions to test these assumptions and include expected or desired changes of a measure in our models. This can even be beneficial beyond reflective measurement models, where DAGs can help to clearly communicate and test different conceptualizations of psychological constructs. In the spirit of Study 3, making our assumptions explicit lets us derive which inferences are warranted and which need additional considerations.

7.3 A Note on Recent Criticism Against Measurement Invariance

In recent years, some scholars have not only voiced their concern about the current practice of investigating MI but also about the concept of MI and its necessity altogether. Most notable in this regard are the articles by Funder and Gardiner (2024), Robitzsch and Lüdtke (2023), and Welzel et al. (2023) (in the following, these three groups of authors are referred to as FRW). FRW question whether MI is really a necessary condition for meaningful latent mean comparisons. They approach their criticisms from a statistical (Robitzsch & Lüdtke, 2023), a conceptual (Funder & Gardiner, 2024), and a somewhat combined standpoint (Welzel et al., 2023). I think it is important to address this criticism here. The manuscripts of this thesis were not only published during the same period as these critical articles but they also take a different standpoint: I strongly argue that measurement invariance *is* important for meaningful latent mean comparisons.

FRW raise important and valid points of criticism. My goal is not to refute every argument raised by FRW. Rather, I want to broadly summarize FRW's points, show where I agree and disagree, and provide a different perspective from a (mostly) methodological standpoint. Fischer et al. (2023) and especially Meuleman et al. (2023) have commented on Welzel et al. (2023), already addressing most of their points. Fischer and Rudnev (2024) have recently published a comment on Funder and Gardiner (2024) in which they provide a comprehensive counter perspective, highlighting the importance of various forms of invariance for personality science. Adding to these

detailed commentaries, I focus on some of FRW's key arguments. Very broadly, I categorize the arguments raised by FRW into three groups: arguments with which I personally (1) agree, (2) partially agree, and (3) disagree.

7.3.1 Agreement

FRW voice valid concern about some issues of investigating MI that are surely not ideal, both from a conceptual and a methodological standpoint. First, Robitzsch and Lüdtke (2023) and Welzel et al. (2023) criticize that trying to establish partial MI (i.e., the invariance of some, but not all parameters) is a threat to (external) validity. I fully agree. Under partial MI, some items are allowed to have group-specific parameters, which corresponds to a "downweighting" of these items in their contribution to the scaling process (Kreiner & Christensen, 2014; Robitzsch & Lüdtke, 2022). This is especially problematic in cases where more than two groups are compared because the partial invariance approach might result in different sets of invariant items that are used for different group comparisons. Even further, both the choice of anchor items that are fixed across groups and the number of invariant items across groups needed for meaningful comparisons are neither trivial nor clear (Pohl et al., 2021; Steenkamp & Baumgartner, 1998). In my opinion, investigating the causes of non-invariance (Sterner et al., 2024) or analyzing the data on the individual item level (e.g., Seifert et al., 2024) are (among others) better options than aiming for partial MI, in cases where MI does not hold between (all) groups.[22]

Second, Funder and Gardiner (2024) and Robitzsch and Lüdtke (2023) denounce the cut-offs of fit indices at which we would state that two parameters are non-invariant (e.g., a decrease in CFI of more than 0.01). They argue that these cut-offs are arbitrary, all derived from simulation studies, and hinge on the data and model conditions used in these simulation studies. This is true and I agree, with some remarks. Almost all practical recommendations for content-related decisions based on statistical analyses are

[22]It has been argued that partial invariance models are very similar to the analysis of *differential item functioning* (DIF; Holland & Wainer, 2012) in an item response theory framework (Thissen, 2024). While the investigations of DIF and MI are conceptually closely related, they have different historical origins (Thissen, 2024). MI-analyses concerned mostly the relations between observed and latent variables, focusing more on influences on the factor structure and the validity of a factor analysis model. DIF-analyses originated later in educational testing to increase fairness across test-takers. My argument against partial MI models is not meant to be extended to DIF-analyses, which are from a different paradigm and beyond the scope of this thesis.

derived from simulation studies. And while Robitzsch and Lüdtke (2023) are correct in saying that assumptions made in simulation studies are often violated in empirical data (thus, making their results less generalizable), simulation studies are still a powerful tool to identify ideal conditions for specific methods and to quantify the impact of violated assumptions on substantive analyses. Nonetheless, Funder and Gardiner (2024) and Robitzsch and Lüdtke (2023) make the important point to never take cut-offs from simulation studies at face value (see also Goretzko et al., 2023). As discussed earlier, it might be helpful to develop dynamic cut-offs for violations of MI, similar to the newly developed tailored cut-offs for single-group CFAs or SEMs (Groskurth et al., 2022; McNeish & Wolf, 2023; Schmalbach et al., 2019). These dynamic cut-offs take parameters that influence the model fit and its evaluation (e.g., model complexity) into account and calculate cut-offs that are ideal for the model and data situation at hand. Additionally, as Funder and Gardiner (2024) suggest, guidance is needed so that researchers can show that non-invariance across groups is meaningfully large enough to distort inference — instead of merely showing that differences in parameter estimates are significantly non-zero. For this, again, simulation studies are needed that investigate "how much non-invariance is too much", ideally under many different model and data conditions.

7.3.2 Partial Agreement

The second category are arguments with which I partially agree. These arguments raise important questions regarding the use of psychological scales but, in my opinion, fall short of an explanation why they are an argument *against* the necessity of MI. All arguments by FRW in this category can very broadly be summarized as follows: MI, whether partial or full, contains no information about the (external) validity of a scale, which FRW say is the more important aspect when comparing constructs across groups. FRW unanimously raise this point, each of them with their own focus.

7.3.2.1 The case for validity: Funder & Gardiner (2024).

First, Funder and Gardiner (2024) criticize that methods to investigate MI are internal to a measurement instrument. These methods provide no information about external validity, that is, the relation of the construct that is measured to other relevant constructs and criteria. Funder and Gardiner (2024) conclude that this external validity

is a better metric of evaluation than MI when the goal is to make comparisons across groups. They make their point by stating that the World Happiness Report rankings by the Gallup World Poll works with only a single-item measure. Thus, it cannot be assessed whether this measure is invariant across countries because common methods to investigate MI draw on inter-item relations (i.e., covariance matrices). Nonetheless, the World Happiness Report is said to be valid, as it shows high levels of external validity. Correlations with other well-being scales have been shown (i.e., convergent validity) and Funder and Gardiner (2024) speak of meaningful correlations of the measure on the country-level.

In general, I agree with Funder and Gardiner (2024) that MI might not be *sufficient* for meaningful group comparisons. External validity, that is, a link to a nomological net of related constructs and criteria is also important to allow for meaningful comparisons. However, in my opinion, it is not a question of "either-or". MI is still *necessary* in that we can be certain that constructs are measured equivalently and external relations, like regression parameters in structural models, are comparable across groups. For example, covariances between latent variables are only comparable if loadings are invariant across groups (i.e., if metric MI holds; Fischer & Rudnev, 2024). Surveys like the World Happiness Report might be a useful and easily understandable tool to broadly assess happiness across many countries. However, using a single-item rating at only one time point to measure happiness does not solve all measurement-related issues, like reliability or MI — these issues just become more difficult or even impossible to address. Thus, whether happiness "scores" are truly comparable cannot be stated with certainty. As Funder and Gardiner (2024) noted, its comparability cannot even be assessed in a factor-analytic framework to investigate MI. As a consequence, comparisons across groups should be made with caution. For example, assume we found a positive correlation of the happiness item with general health in two countries. This would be a sign of validity, assuming that happiness is associated with general health (e.g., Steptoe, 2019). However, if the correlation was larger in one country than in the other country, MI would be a necessary condition to be able to interpret the difference of the correlation between the two countries. We can only substantively interpret these differences in correlations if we can be certain that it was really happiness that we measured in both

countries (and not something closely related but slightly different).[23]

To increase the external validity of psychological measures, Funder and Gardiner (2024) make the important suggestion to move more toward the collection and use of behavioral data, instead of mainly self-report data. This, in my opinion, is an excellent suggestion to enrich psychological data by more than just questionnaire data which rely on the assumption of MI. Beyond the examples by Funder and Gardiner (2024), smartphone sensing, that is, the objective collection of smartphone usage behavior like screen time, are one possible option of these new behavioral data sources (Stachl, Pargent, et al., 2020; Stachl, Au, et al., 2020). These data can even be combined with psychological questionnaires in the form of self-reports (Reiter & Schoedel, 2023; Schoedel et al., 2020).

7.3.2.2 The case for validity: Robitzsch and Lüdtke (2023).

Second, Robitzsch and Lüdtke (2023) suggest that the requirement of MI might even pose a threat to the validity of a scale; specifically in cases where a differential functioning of items across groups is imposed by the definition of a construct. Removing items whose parameters are non-invariant would thus decrease the validity of a scale if the item is or, by definition of the construct, should be non-invariant. Instead, items should be added to or removed from a scale on content-related and non-statistical grounds. Non-invariance should only be reported as a further source of uncertainty in parameter estimates. For meaningful group comparisons, however, the focus should be on (external) validity of the scale.

I agree with Robitzsch and Lüdtke (2023) that the removal of items from a scale should never be based solely on investigations of MI. In fact, the main idea behind the causal framework in Study 3 is that measures of a construct might function differently across groups. Theoretical considerations regarding this differential functioning should be incorporated in measurement models to account for item-by-group interactions (i.e., non-invariance). For example, as demonstrated in Study 3, the effects of covariates on the parameters of a measurement model could be included in the model to account for these interactions. Even further, I would differentiate between the items measuring a

[23] Even if the correlations between happiness and a criterion were equal in both countries, this would not necessarily be evidence that happiness was measured equivalently (cf. Fischer & Rudnev, 2024).

construct and the measurement model relating these items to a latent variable which represents the construct. When investigating MI, what we are really investigating is whether the measurement model we assume to hold is tenable across groups. If this test of MI fails, this does not necessarily entail that the items are inadequate to measure the construct. Instead, it could just be that the measurement model needs improvement, in that it might be too simple or not incorporating important relations between the items and covariates. Thus, in my opinion, the argument by Robitzsch and Lüdtke (2023) — that item-by-group interactions or non-invariance might be inherent in a construct by its definition — is not an argument against the necessity of MI. Rather, it underscores the importance to think outside the measurement model and include potential interaction effects on item parameters in our models when investigating MI (see Study 3).

In general, I would again question whether shifting the focus from MI to external validity, thereby labeling MI as unnecessary, will make latent mean comparisons more meaningful. Instead, investigations of external validity should be added to assessments of MI. Unfortunately, Robitzsch and Lüdtke (2023) neglected to provide methodological guidance on how applied researchers can proceed in this regard. In their defense, the article by Robitzsch and Lüdtke (2023) focused on methodological rigor and proofs, so an applied demonstration of alternative approaches was beyond the scope of their work. Nonetheless, the main issue I see with the argument by Robitzsch and Lüdtke (2023) is that, without any alternative to investigations of MI, a key insight for readers might be that they can simply ignore MI and proceed with their analyses as usual. For example, He et al. (2024) assessed the invariance of their measures, found that (scalar) MI is not given, and proceeded with their analyses because Robitzsch and Lüdtke (2023) showed that MI "is not a prerequisite for meaningful and valid group comparisons" (p. 4). I see this as problematic because I would question whether "we do not have to investigate MI" is the opinion of Robitzsch and Lüdtke (2023). Rather, I understand their point to be that meaningful group comparisons need different or additional justifications than only the establishment of MI. Investigating MI and using Robitzsch and Lüdtke (2023) as a citation in case it is not given to proceed with the analyses, is, in my opinion, not a productive way forward for latent mean comparisons.

7.3.2.3 The case for validity: Welzel et al. (2023).

Finally, Welzel et al. (2023) criticize that measures of a construct are too easily delegitimized if they are not invariant. Especially if the measures are "strongly reality-linked" (p. 1374), that is, if external validity is given by means of correlations with other constructs or criteria, this be evidence of comparability of said measure.

I agree with Welzel et al. (2023) that a measure of a construct should not be discarded solely on the ground that it is not invariant. Particularly not, if this non-invariance was only found by MG-CFA, which Welzel et al. (2023) state to be the primary method of investigation. MI should always be investigated in more detail by means of advanced methods (e.g., De Roover et al., 2022; Sterner & Goretzko, 2023) or by reasoning about causes of non-invariance (Sterner et al., 2024). I want to highlight again that I see non-invariance as an important finding by itself, and that it should always be treated as such (see also Maassen et al., 2023). Fischer and Rudnev (2024) emphasised the importance of invariance that even goes beyond the statistical aspects of MI; for example regarding the conceptualization or operationalization of constructs. In general, non-invariant measures tell us something about how and why groups interpret a construct differently (Putnick & Bornstein, 2016).

Meuleman et al. (2023) have already refuted the claim by Welzel et al. (2023) that external validity of a measure is evidence for the comparability of its scores across groups. To this I add that — similar to the argument on the World Happiness Report by Funder and Gardiner (2024) — calling measures "comparable" is, initially, a linugistic matter. Of course, weaker definitions of comparability than the one postulated by MI are possible and might also be reasonable. However, every definition of comparability is obligated to define (and demonstrate) which inferences are warranted under which conditions. For MI, clearly defined and testable conditions exist that are directly related to the statistical inferences they are licensing (e.g., Vandenberg & Lance, 2000). External validity alone is not a sufficient condition to extend our content-related claims to the construct level, let alone compare measurements of the construct across groups. The fact remains: every difference we observe between groups can only be interpreted as a true difference if it cannot be attributed to differences in measurement. Thus, comparability of measurements across groups hinges on invariant measurements

(according to its definition in the context of MI; Meuleman et al., 2023).

7.3.3 Disagreement

The third category are arguments with which I personally disagree. In my opinion, they to some extent provide a biased view on the methodological literature and neglect important recent developments. My goal is to clear up potential misunderstandings between the applied and the methodological literature on MI. I do not want to deny questionable practices or personal experiences that the authors witnessed and reported in their respective articles.

First, Funder and Gardiner (2024) insinuate a harsh language in the methodological literature. They criticize the "prohibitionist tone" (p. 2) that is apparently struck in discussions about MI. They argue that when researchers speak of "failure [to establish MI]" or of a "violation [of MI]", these terms imply that in these cases cross-cultural data cannot be taken seriously or that the data should be ignored. However, words like "failure" or "violation" are neither prohibitionist nor unique to the topic of MI. They are common statistical terms. We speak of "failure to reject a null hypothesis" or "violations of statistical assumptions" in almost every statistical analysis. The terms are thus scientific jargon and have no hidden meaning. Welzel et al. (2023) call non-invariance (as found by MG-CFA) a "lethal" (p. 1370) and later a "fatal verdict that delegitimizes the further use of the respective construct in cross-cultural comparison" (p. 1372). Maybe the scientific community surrounding cross-cultural research is more severe in their tone when it comes to discussions about MI. But from my personal reading of the methodological literature, none of these drastic wordings are justified. In fact, the very reason why the many different methods to investigate MI presented in this thesis and elsewhere (e.g., Kim et al., 2017) were developed, is to enable researchers to analyze their data even in cases of non-invariance; or to at least properly diagnose its sources and causes. Maassen et al. (2023) explicitly state that non-invariance should not be seen as a "roadblock" (p. 12). Instead, it can be used to deepen our understanding of constructs in that it might tell us something about how and why different groups interpret a construct in different ways. The causal framework in Study 3 propagates and supports this more constructive view on investigations of MI.

Second, Funder and Gardiner (2024) imply that their interpretation of the literature on MI is that data cannot be used if the measurement models of the constructs are found to be non-invariant. They see this as particularly problematic in the context of cross-cultural research because it is difficult to establish MI across many countries. As just mentioned, I would opt for a different reading of the methodological literature: ignoring data when models are non-invariant is not the general opinion (e.g., Maassen et al., 2023). There are a wide variety of methods that were developed to specifically cater to situations where many groups have to be compared, for example in cross-cultural research. The methods presented by Kim et al. (2017) as well as MMG-EFA (De Roover, 2021; De Roover et al., 2022) or EFA trees (Sterner & Goretzko, 2023) are examples of methods that enable investigations of MI in cases of many groups. They are *not* developed to reach a binary decision regarding (non-)invariance but to enable meaningful data analysis in cases where exact MI does not hold between all groups.

Third, Welzel et al. (2023) base most of their methodological criticism on a case against MG-CFA. Meuleman et al. (2023) have already commented on their methodological claims in great detail. Just like Funder and Gardiner (2024), Welzel et al. (2023) missed the opportunity to acknowledge the large number of studies on advanced methods to investigate MI that were developed in the last 15 to 20 years. Many of their arguments depend on their premise that MG-CFA is an "increasingly fashionable methodology" used in a "newly spreading type of study" (p. 1370). Not only is MG-CFA one of the oldest methods to investigate MI, it is also (unfortunately) not of much popularity in empirical studies (Maassen et al., 2023). From the perspective of causal inference, we also argued that MG-CFA could be seen as the method with the strictest assumptions among all methods to investigate MI (see Study 3). For this reason, too, an acknowledgement of the great variety of MI-methods might help to provide a more constructive view on MI. Most of the methodological flaws regarding investigations of MI that Welzel et al. (2023) condemn are thus unique to MG-CFA. These flaws have already been heard and addressed by the many colleagues that have developed more advanced methods.

7.3.4 Conclusion

In summary, I think it is important that methodological concepts like MI are being scrutinized from different perspectives. Through critical discussion, we can improve our understanding and implementation of MI in psychological science. However, it is equally important that these discussions take into account all available information. Most notably, criticism regarding the status quo should be accompanied by alternative solutions. As always in statistical analyses, there is no free lunch: to enable meaningful conclusions, we have to make assumptions and we have to make them explicit. Group comparisons with large samples, validated measures, and plausible models can be meaningful, even when viewed against the background of violated assumptions, like non-invariance. But this lack of MI should be properly assessed with appropriate methods and transparently reported so that readers themselves can evaluate the validity of the claims in a study. Above all, non-invariance does not preclude the analysis of any data — it should instead be treated as an important finding by itself.

7.4 General Conclusion

Most of psychology relies on measurements of constructs produced by questionnaires. This concerns both psychological research and psychological assessment, for example, in therapy, organizations, or schools. To provide comparable measurements, regardless of a person's background variables, we have to ensure that our measurements are invariant across groups or time. The aim of this thesis was to extend the vast literature of methodological research on measurement invariance. It introduced a new method to investigate MI among multiple covariates, provided an overview and comparison of EFA-based approaches, and suggested a causal framework for future research. These contributions hopefully motivate applied researchers to more frequently consider MI in empirical studies that compare psychological constructs. At the same time, the studies of this thesis have the potential to pave the way for future research on methodological challenges, for example a conceptualization of longitudinal MI.

To close, psychology can only benefit from a more thorough treatment of the invariance of its measurements. Only then can we as psychologists be sure that we are not comparing apples and extraversion, or intelligence and oranges.

7.5 References

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D. (2006b). When Does Measurement Invariance Matter? *Medical Care*, *44*(11), 176–181. https://doi.org/10.1097/01.mlr.0000245143.08679.cc

Borsboom, D., Deserno, M. K., Rhemtulla, M., Epskamp, S., Fried, E. I., McNally, R. J., ... & Waldorp, L. J. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers*, *1*, 58. https://doi.org/10.1038/s43586-021-00055-w

Cao, C., & Liang, X. (2022a). Sensitivity of Fit Measures to Lack of Measurement Invariance in Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(2), 248–258. https://doi.org/10.1080/10705511.2021.1975287

Cao, C., & Liang, X. (2022b). The Impact of Model Size on the Sensitivity of Fit Measures in Measurement Invariance Testing. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(5), 744–754. https://doi.org/10.1080/10705511.2022.2056893

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504.

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

De Roover, K. (2021). Finding Clusters of Groups with Measurement Invariance: Unraveling Intercept Non-Invariance with Mixture Multigroup Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 663–683. https://doi.org/10.1080/10705511.2020.1866577

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2022). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups.

Psychological Methods, *27*, 281–306. https://doi.org/10.1037/met0000355

Donker, T., Griffiths, K. M., Cuijpers, P., & Christensen, H. (2009). Psychoeducation for depression, anxiety and psychological distress: A meta-analysis. *BMC Medicine*, *7*(1), 79. https://doi.org/10.1186/1741-7015-7-79

Fischer, R., Karl, J. A., Fontaine, J. R. J., & Poortinga, Y. H. (2023). Evidence of Validity Does not Rule out Systematic Bias: A Commentary on Nomological Noise and Cross-Cultural Invariance. *Sociological Methods & Research*, *52*(3), 1420–1437. https://doi.org/10.1177/00491241221091756

Fischer, R., & Rudnev, M. (2024). From MIsgivings to MIse-en-scène: the role of invariance in personality science. *European Journal of Personality*

Fokkema, M., Smits, N., Kelderman, H., & Cuijpers, P. (2013). Response shifts in mental health interventions: An illustration of longitudinal measurement invariance. *Psychological Assessment*, *25*(2), 520–531. https://doi.org/10.1037/a0031669

Fried, E. I., Borkulo, C. D. van, Epskamp, S., Schoevers, R. A., Tuerlinckx, F., & Borsboom, D. (2016). Measuring depression over time . . . Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. *Psychological Assessment*, *28*(11), 1354–1367. https://doi.org/10.1037/pas0000275

Fried, E. I., Epskamp, S., Nesse, R. M., Tuerlinckx, F., & Borsboom, D. (2016). What are'good'depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. *Journal of affective disorders*, *189*, 314–320. https://doi.org/10.1016/j.jad.2015.09.005

Fried, E. I., von Stockert S., Haslbeck J. M. B., Lamers F., Schoevers R. A., & Penninx B. W. J. H. (2020). Using network analysis to examine links between individual depressive symptoms, inflammatory markers, and covariates. *Psychological Medicine*, *50*, 2682–2690. https://doi.org/10.1017/S0033291719002770

Funder, D. C., & Gardiner, G. (2024). MIsgivings about measurement invariance. *European Journal of Personality*, 08902070241228338. https://doi.org/10.1177/

08902070241228338

Goretzko, D., Siemund, K., & Sterner, P. (2023). Evaluating Model Fit of Measurement Models in Confirmatory Factor Analysis. *Educational and Psychological Measurement*, *84*(1), 123–144. https://doi.org/10.1177/00131644231163813

Groskurth, K., Bhaktha, N., & Lechner, C. (2022). *Making model judgments ROC(K)-solid: Tailored cutoffs for fit indices through simulation and ROC analysis in structural equation modeling*. OSF. https://doi.org/10.31234/osf.io/62j89

Gunn, H. J., Grimm, K. J., & Edwards, M. C. (2020). Evaluation of Six Effect Size Measures of Measurement Non-Invariance for Continuous Outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 503–514. https://doi.org/10.1080/10705511.2019.1689507

He, J., Cui, S., Cui, T., Barnhart, W. R., Han, J., Xu, Y., & Nagata, J. M. (2024). Exploring the associations between muscularity teasing and eating and body image disturbances in Chinese men and women. *Body Image*, *49*, 101697. https://doi.org/10.1016/j.bodyim.2024.101697

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Heene, M., Hilbert, S., Freudenthaler, H. H., & Bühner, M. (2012). Sensitivity of SEM Fit Indexes With Respect to Violations of Uncorrelated Errors. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 36–50. https://doi.org/10.1080/10705511.2012.634710

Holland, P. W., Wainer, H. (2012). *Differential item functioning*. Routledge.

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 524–544. https://doi.org/10.1080/10705511.2017.1304822

Kolbe, L., Molenaar, D., Jak, S., & Jorgensen, T. D. (2022). Assessing measure-

ment invariance with moderated nonlinear factor analysis using the R package OpenMx. *Psychological Methods*, *Advance Online Publication*. https://doi.org/10.1037/met0000501

Kreiner, S., & Christensen, K.B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, *79*, 210–231. https://doi.org/10.1007/s11336-013-9347-z

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods*, *28*(4), 905–924. https://doi.org/10.1037/met0000441

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

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

McNeish, D. (2023). Dynamic fit index cutoffs for categorical factor analysis with Likert-type, ordinal, or binary responses. *American Psychologist*, *78*(9), 1061–1075. https://doi.org/10.1037/amp0001213

McNeish, D., & Wolf, M. G. (2023). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, *28*(1), 61–88. https://doi.org/10.1037/met0000425

McNeish, D., & Wolf, M. G. (2024). Direct Discrepancy Dynamic Fit Index Cutoffs for Arbitrary Covariance Structure Models. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–28. https://doi.org/10.1080/10705511.2024.2308005

Meuleman, B., Żółtak, T., Pokropek, A., Davidov, E., Muthén, B., Oberski, D. L.,

Billiet, J., & Schmidt, P. (2023). Why Measurement Invariance is Important in Comparative Research. A Response to Welzel et al. (2021). *Sociological Methods & Research*, *52*(3), 1401–1419. https://doi.org/10.1177/00491241221091755

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. https://doi.org/10.1037/a0022955

Oberski, D. L. (2014). Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis*, *22*(1), 45–60.

Oberski, D. L., Vermunt, J. K., & Moors, G. B. D. (2015). Evaluating Measurement Invariance in Categorical Data Latent Variable Models with the EPC-Interest. *Political Analysis*, *23*(4), 550–563. https://www.jstor.org/stable/24573192

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best Practices in Supervised Machine Learning: A Tutorial for Psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162559. https://doi.org/10.1177/25152459231162559

Partsch, M., Sterner, P., & Goretzko, D. (2024). *A Simulation Study on the Interaction Effects of Underfactoring and Nuisance Parameters on Model Fit Indices*. OSF. https://doi.org/10.31234/osf.io/qy2e3

Pohl, S., Schulze, D., & Stets, E. (2021). Partial Measurement Invariance: Extending and Evaluating the Cluster Approach for Identifying Anchor Items. *Applied Psychological Measurement*, *45*(7-8), 477–493. https://doi.org/10.1177/01466216211042809

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

Reiter, T., & Schoedel, R. (2023). Never miss a beep: Using mobile sensing to investigate (non-)compliance in experience sampling studies. *Behavior Research*

Methods. https://doi.org/10.3758/s13428-023-02252-9

Robitzsch, A., & Lüdtke, O. (2022). Mean Comparisons of Many Groups in the Presence of DIF: An Evaluation of Linking and Concurrent Scaling Approaches. *Journal of Educational and Behavioral Statistics*, *47*(1), 36–68. https://doi.org/10.3102/10769986211017479

Robitzsch, A., & Lüdtke, O. (2023). Why Full, Partial, or Approximate Measurement Invariance Are Not a Prerequisite for Meaningful and Valid Group Comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 859–870. https://doi.org/10.1080/10705511.2023.2191292

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*, 31–57. https://doi.org/10.1177/0013164413498257

Savalei, V. (2012). The Relationship Between Root Mean Square Error of Approximation and Model Misspecification in Confirmatory Factor Analysis Models. *Educational and Psychological Measurement*, *72*(6), 910–932. https://doi.org/10.1177/0013164412452564

Savalei, V., Brace, J. C., & Fouladi, R. T. (2023). We need to change how we compute RMSEA for nested model comparisons in structural equation modeling. *Psychological Methods*, No Pagination Specified–No Pagination Specified. https://doi.org/10.1037/met0000537

Schmalbach, B., Irmer, J., & Schultze, M. (2019). ezCutoffs: Fit measure cutoffs in SEM. *R Package Version*, *1*(1).

Schoedel, R., Pargent, F., Au, Q., Völkel, S. T., Schuwerk, T., Bühner, M., & Stachl, C. (2020). To Challenge the Morning Lark and the Night Owl: Using Smartphone Sensing Data to Investigate Day–Night Behaviour Patterns. *European Journal of Personality*, *34*(5), 733–752. https://doi.org/10.1002/per.2258

Seifert, I. S., Rohrer, J. M., & Schmukle, S. C. (2024). Using within-person change in three large panel studies to estimate personality age trajectories. *Journal of Personality and Social Psychology*, *126*(1), 150–174. https://doi.org/10.1037/

pspp0000482

Sijtsma, K. (2006). Psychometrics in Psychological Research: Role Model or Partner in Science? *Psychometrika*, *71*(3), 451. https://doi.org/10.1007/s11336-006-1497-9

Sijtsma, K. (2012). Future of Psychometrics: Ask What Psychometrics Can Do for Psychology. *Psychometrika*, *77*(1), 4–20. https://doi.org/10.1007/s11336-011-9242-4

Somaraju, A. V., Nye, C. D., & Olenick, J. (2022). A Review of Measurement Equivalence in Organizational Research: What's Old, What's New, What's Next? *Organizational Research Methods*, *25*(4), 741–785. https://doi.org/10.1177/10944281211056524

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. https://doi.org/10.1073/pnas.1920484117

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality Research and Assessment in the Era of Machine Learning. *European Journal of Personality*, *34*(5), 613–631. https://doi.org/10.1002/per.2257

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

Steptoe, A. (2019). Happiness and health. *Annual Review of Public Health*, *40*(1), 339–359. https://doi.org/10.1146/annurev-publhealth-040218-044150

Sterner, P., & Goretzko, D. (2023). Exploratory Factor Analysis Trees: Evaluating Measurement Invariance Between Multiple Covariates. *Structural Equation Modeling: A Multidisciplinary Journal*, *30:6*, 871–886. https://doi.org/10.1080/10705511.2023.2188573

Sterner, P., Pargent, F., Deffner, D., & Goretzko, D. (2024). A Causal Framework for

the Comparability of Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12. https://doi.org/10.1080/10705511.2024.2339396

Thissen, D. (2024). A Review of Some of the History of Factorial Invariance and Differential Item Functioning. *Multivariate Behavioral Research*, 1–25. https://doi.org/10.1080/00273171.2024.2396148

Tursi, M. F. de S., Baes, C. von W., Camacho, F. R. de B., Tofoli, S. M. de C., & Juruena, M. F. (2013). Effectiveness of psychoeducation for depression: A systematic review. *Australian & New Zealand Journal of Psychiatry*, *47*(11), 1019–1031. https://doi.org/10.1177/0004867413491154

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

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Vandenberg, R. J. (2002). Toward a Further Understanding of and Improvement in Measurement Invariance Methods and Procedures. *Organizational Research Methods*, *5*(2), 139–158. https://doi.org/10.1177/1094428102005002001

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70.

Welzel, C., Brunkert, L., Kruse, S., & Inglehart, R. F. (2023). Non-invariance? An Overstated Problem With Misconceived Causes. *Sociological Methods & Research*, *52*(3), 1368–1400. https://doi.org/10.1177/0049124121995521

World Health Organization. (2022). *ICD-11: International Classification of Diseases (11th Revision)*. https://icd.who.int/