Novel approaches to assess measurement invariance

Kolbe, L.

[Link to publication](#)

Citation for published version (APA):
Kolbe, L. (2022). *Novel approaches to assess measurement invariance*.

Novel Approaches to Assess Measurement Invariance

Laura Kolbe

Laura Kolbe

Novel Approaches to Assess Measurement Invariance

Laura Kolbe

Novel Approaches to Assess Measurement Invariance

Academisch proefschrift

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. P.P.C.C. Verbeek

ten overstaan van een door het College voor Promoties ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op maandag 10 oktober 2022, te 14.00 uur

door

Laura Kolbe

geboren te Zwolle

Promotiecommissie

Promotor:	Prof. dr. F.J. Oort	Universiteit van Amsterdam
Copromotores:	Dr. T.D. Jorgensen	Universiteit van Amsterdam
	Dr. S.J. Jak	Universiteit van Amsterdam
Overige leden:	Prof. dr. M.E. Timmerman	Rijksuniversiteit Groningen
	Prof. dr. D. Borsboom	Universiteit van Amsterdam
	Dr. K.J. Kan	Universiteit van Amsterdam
	Dr. M.G.E. Verdam	Universiteit Leiden
	Prof. dr. L.A. van der Ark	Universiteit van Amsterdam

Faculteit der Maatschappij- en Gedragswetenschappen

Contents

Chapter 1

General Introduction

1.1   Measuring Psychological Constructs

The measurement of psychological constructs is at the heart of social and behavioral sciences. Psychological constructs are often referred to as latent constructs because they are not directly observable. As a result, these constructs are measured with observable variables such as responses to questionnaire or test items that are assumed to represent the latent construct. Whether a researcher wants to study social anxiety or an educator wants to compare students on numerical ability, research and practice often depend on measures of latent constructs. It is even fair to say that the measurement of latent constructs can have great impact on individual lives and society in general. For example, a social anxiety intervention can be categorized as effective or not based on observed measures of social anxiety in a sample of adolescents, or students may or may not receive additional support on numerical ability depending on their observed score on a numerical ability test. Given the dominant role of such tests and questionnaires in social and behavioral sciences and applications (Brennan, 2004), it is important that the items on these measurement instruments function the same way across individuals or groups. This requirement is also referred to as measurement invariance.

There has been a growing awareness regarding the importance of measurement invariance (see Drasgow, 1984; Jöreskog, 1971; Vandenberg & Lance, 2000; Cheung & Rensvold, 1999; Steenkamp & Baumgartner, 1998; Meredith, 1993; Little, 1997; Van de Vijver & Fischer, 2009; Milfont & Fischer, 2010). If the assumption of measurement invariance is violated, the observed scores on a test or questionnaire not only depend on the latent construct intended to be measured but also on variables other than the latent construct. As a result, individuals with the same level of the latent construct may have different expected observed scores. This is problematic because researchers or practitioners might conclude that groups or individuals differ on the latent construct when the differences in observed scores actually arise from differential measurement caused by variables other than the latent construct. Hence, measurement invariance is a necessary condition in order to meaningfully compare groups or individuals on latent constructs and should be assessed before making such comparisons.

A common class of methods for assessing measurement invariance is confirmatory factor analysis (CFA), which is a family of statistical analyses within the structural equation modeling (SEM) framework. A CFA model describes the dependencies between a set of observed variables by using a limited number of so-called common factors. These common factors represent the latent constructs that are measured by the observed variables. Measurement invariance can be assessed in CFA models by means of a comparison of specific features of the model across different levels of variables other than the latent construct. These other variables are referred to as background variables in this dissertation, but are also commonly called violators (see Barendse et al., 2010) or covariates (see Bauer, 2017). The next paragraph shortly addresses different levels of measurement invariance in the context of CFA.

1.2 Measurement Invariance

Consider a set of observed indicators X (e.g., items) measuring the latent construct of interest T, and a set of background variables V (e.g., age or gender). The definition of measurement invariance can be mathematically expressed as

$$f_1(X|T,V) = f_2(X|T) \tag{1.1}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ denote the conditional probability distributions of observed indicators X. The mathematical expression states that measurement invariance holds if the distribution of X depends only on the latent construct T and is invariant with respect to background variables V (Mellenbergh, 1989). If measurement invariance does not hold (i.e., $f_1 \neq f_2$), the distribution of the observed indicators depends not only on the latent construct T but also directly on background variables V. Hence, with a lack of measurement invariance, individuals with an equal standing on the latent construct may have different expected values of X, and differences in the observed scores may not imply true differences in T. Violations of measurement invariance are also referred to as differential item functioning (DIF), measurement bias, or measurement noninvariance. A distinction can be made between full and partial invariance, both indicating different degrees of measurement invariance. Full invariance implies that all observed indicators (e.g., all items of a test or questionnaire) are measurement invariant, whereas partial invariance implies that measurement invariance only holds for a subset of the observed indicators.

In addition to different degrees of measurement invariance, different levels of measurement invariance have been defined (Meredith, 1993; Steenkamp & Baumgartner, 1998; Horn & McArdle, 1992). Consider a CFA model in which the continuous observed variables X serve as indicators of the latent construct modeled as common factor T. This model can be specified as

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\lambda} t_i + \boldsymbol{\varepsilon}_i \tag{1.2}$$

where $\boldsymbol{x}_i$ is a vector of observed indicator scores for individual i, $\boldsymbol{\tau}$ is a vector of indicator intercepts, and $\boldsymbol{\lambda}$ is a vector of factor loadings. Moreover, t_i is the common-factor score and $\boldsymbol{\varepsilon}_i$ is a vector of residual scores with variances $\boldsymbol{\theta}$. The different levels of measurement invariance can be ordered from least to most restrictive. The least restrictive level of measurement invariance is called configural invariance. Configural invariance implies equal factor loading patterns across the background variable, that is, the latent construct is being measured by the same indicators across all levels of the background variable. The next more restrictive level of measurement invariance is metric invariance, which additionally implies equal factor loadings across the background variable. A yet more restrictive level of measurement invariance is scalar invariance, which in addition to equal factor loadings implies equal indicator intercepts across the background variable. The most restrictive level of measurement invariance is called strict invariance, additionally implying equal residual variances across the background variable.

One of the traditional methods to evaluate these levels of measurement invariance across a categorical background variable (e.g., group membership) is multiple-group CFA (MGCFA; Vandenberg & Lance, 2000). In MGCFA, the data are divided into two or more independent groups based on V and a CFA model, as shown in Equation 1.2, is estimated for each group separately. Measurement invariance can then be assessed by comparing the fit of models with and without increasingly restrictive equality constraints on the measurement parameters (e.g., the factor loadings $\boldsymbol{\lambda}$ or intercepts $\boldsymbol{\tau}$) across the background variable. Full invariance can be examined with an omnibus test for a particular level of measurement invariance for all indicators simultaneously (Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Finch & French, 2018; Marsh, 1994). When the omnibus null hypothesis of full invariance is rejected, one can proceed with examining partial invariance. Under partial invariance, groups or individuals can still be validly compared on the latent construct as long as violations of measurement invariance are correctly detected and modeled. Establishing partial invariance requires assessing each indicator separately for measurement invariance by comparing the fit of a model with and without equality constraints on that indicator's parameters. This way, each indicator can be assessed individually while holding a subset of other indicators invariant across the background variable. These latter indicators are also called anchor indicators and are used to link the metric of the common factors across the background variable when assessing measurement invariance on indicator-level. Anchor indicators can be selected using an anchor-selection strategy (for an overview, see Kopf et al., 2015a).

As MGCFA relies on splitting the data into two groups, it is best suited for categorical background variables. Alternative methods for assessing measurement invariance have been proposed, among which are restricted factor analysis (RFA; Oort, 1992), multiple indicator multiple cause (MIMIC; Jöreskog & Goldberger, 1975), and moderated nonlinear factor analysis (MNLFA; Bauer & Hussong, 2009) models. In contrast to MGCFA, these methods aggregate the data over V (e.g., group membership or age) and will therefore be referred to as single-group methods throughout this dissertation. There are several advantages of single-group methods over the MGCFA method, including a potentially higher statistical power to detect violations of measurement invariance because the number of parameters to be estimated is reduced by aggregating the data over subsamples in all single-group methods (see Barendse et al., 2012). Another advantage of single-group methods over MGCFA is that they easily accommodate tests for measurement invariance with respect to a continuous background variable V. In MGCFA, testing for measurement invariance with respect to a continuous background variable would require categorizing the continuous variable scores, which may lead to a loss of power and measurement reliability (MacCallum et al., 2002). In addition, because the single-group methods do not rely on splitting the data into groups, they accommodate testing for measurement invariance across multiple continuous and categorical background variables simultaneously and allow for more complex functional relationships, such as interactions or curvilinear effects of V.

Various single-group methods, including RFA and MNLFA, have been proposed for

the purpose of assessing measurement invariance. These methods share the same general form of the measurement model, but differ in the way the background variable V is modeled and differences in measurement parameters are estimated. Although some studies have investigated the performance of single-group methods (see Barendse et al., 2010, 2012; Bauer et al., 2020; Woods & Grimm, 2011), more extensive research is needed to further analyze the performance of these methods in different situations. There are multiple challenges and unsolved problems regarding the use of single-group methods for measurement invariance assessment. For example, it is unknown how the performance of these methods can be improved in order for a more valid assessment of measurement invariance, or which method can be preferred over the other in certain situations. The performance of these methods thus remains subject of ongoing research.

1.3   Aim and Outline

This dissertation addresses novel approaches to assess measurement invariance in the framework of SEM. More specifically, the focus of this dissertation is on new ways to assess measurement invariance using single-group methods. The dissertation starts with a study of the single-group method RFA. An RFA model is not readily suited for assessing metric invariance and is therefore commonly extended with latent moderated structural equations (LMS). As LMS is implemented in limited SEM computer programs and does not provide most traditional SEM fit indices, product indicators (PI; Kenny & Judd, 1984) can be used instead in RFA models to assess metric invariance. In **Chapter 2**, the use of PI in RFA models is introduced and illustrated. In order to further investigate the performance of the PI method in RFA, a more extensive simulation study is performed in **Chapter 3**. This study not only includes a comparison between PI and LMS, but also a comparison between two anchor-selection strategies. The anchor-selection strategy that performs best in the first part of the study is used in the second part of the study in which the performance of PI and LMS is compared. In contrast to the MGCFA method, RFA comes with the additional assumptions of equal common-factor variances across different levels of the background variable (i.e., common-factor homoskedasticity) and equal indicators' residual variances across different levels of the background variable (i.e., residual homoskedasticity). The robustness of RFA to violations of common-factor and residual homoskedasticity is relatively unexplored (see Chun et al., 2016; Harpole, 2015, for exceptions). In **Chapter 4**, a study is presented in which the performance of RFA is examined in situations with different magnitudes of common-factor and residual variance differences across the background variable. MNLFA (Bauer & Hussong, 2009; Bauer, 2017) models may be a more suitable alternative to RFA in the presence of common-factor or residual heteroskedasticity, because such models do not require assuming common-factor or residual homoskedasticity with respect to the background variable. MNLFA has not yet been compared to other methods for measurement-invariance assessment and its statistical properties in conditions with small samples and continuous indicators are yet unknown.

Therefore, this chapter focuses on comparing the performance of RFA and MNLFA to test for measurement invariance under common-factor and residual homoskedasticity and heteroskedasticity. **Chapter 5** presents a study on assessing measurement invariance through MNLFA with the R (R Core Team, 2021) package `OpenMx` (Boker et al., 2011). The aim is to make MNLFA more accessible for researchers by providing detailed guidelines on performing the method in this open-source SEM software. The chapter therefore also includes a tutorial. Finally, the dissertation concludes with **Chapter 6**, which summarizes the research findings of this dissertation and raises several additional research questions.

Chapter 2

Using Product Indicators in Restricted Factor Analysis to Assess Measurement Invariance

Abstract

When sample sizes are too small to support multiple-group models, an alternative method to evaluate measurement invariance is restricted factor analysis (RFA), which is statistically equivalent to the more common multiple-indicator multiple-cause (MIMIC) model. Although these methods traditionally were capable of detecting only violations of scalar invariance, RFA can be extended with latent moderated structural equations (LMS) to assess violations of metric invariance. As LMS is implemented in limited structural equation modeling (SEM) software (e.g., M*plus*), we propose the use of product indicators (PI) in RFA models, which are available to use in any SEM software. Using simulated data, we illustrate how this method can be used to assess measurement invariance, and we compare the conclusions with those reached using LMS in M*plus*. Both methods obtain comparable results, indicating that the PI method is a viable alternative to LMS for researchers without access to SEM software featuring LMS.

2.1 Introduction

Measurement invariance entails that scales function similarly across groups, irrespective of true differences in the construct that the scale was designed to measure. Let T denote the construct of interest measured by a set of observed variables X. Moreover, let V be a set of background variables other than T. The formal definition of measurement invariance can be expressed as follows (Mellenbergh, 1989):

$$f_1(X|T = t, V = v) = f_2(X|T = t) \tag{2.1}$$

where $f_1(\cdot)$ is the conditional distribution of X given T and V, and $f_2(\cdot)$ the conditional distribution of X given T. If measurement invariance holds (i.e., $f_1 = f_2$), the measurement of T by X is invariant with respect to V. But if measurement invariance does not hold (i.e., $f_1 \neq f_2$), the measurement of T by X functions differently with respect to V. A violation of measurement invariance is often referred to as differential item functioning (DIF). A distinction can be made between violations of scalar invariance and metric invariance, also referred to as uniform and nonuniform DIF, respectively. Uniform DIF implies that the extent of DIF is constant for all levels of the construct T, whereas nonuniform DIF implies that the extent of DIF varies with T.

A common method to assess measurement invariance with respect to a grouping variable is multiple-group confirmatory factor analysis (MGCFA; Vandenberg & Lance, 2000), which requires sufficiently large samples for each group. An alternative for assessing measurement invariance is restricted factor analysis (RFA; Oort, 1992, 1998). An advantage of this method over MGCFA is that the background variable V may be categorical or continuous, observed or latent, and multiple background variables can be investigated simultaneously. Moreover, RFA does not require the division of the sample into subsamples by V. The latter advantage comes at the cost of additional assumptions—namely, homogeneity of residual variances across groups[1]. If these additional assumptions hold, RFA should have more power than MGCFA to detect violations of measurement invariance.

When using RFA, the background variable V is added to a common factor model as an exogenous variable that covaries with T. Uniform DIF can be assessed by testing the significance of direct effects of V on X. To assess nonuniform DIF, an extension for modeling latent interactions is required. RFA is commonly extended with latent moderated structural equations (LMS; Barendse et al., 2010). This allows for assessing nonuniform DIF by testing the significance of interaction effects of $T \times V$ on X. Although this method generally has high power to detect DIF (Barendse et al., 2010, 2012; Woods & Grimm, 2011), a disadvantage is that LMS is only implemented in the commercial

---

[1]In traditional RFA models, common-factor variances are also assumed to be equal across groups. However, when extending RFA to include a latent interaction factor with product indicators (described immediately following), differences in common-factor variances can be captured by the covariance between the common factor and the latent interaction factor.

structural equation modeling (SEM) software M*plus* (L. K. Muthén & Muthén, 2012)[2]. Moreover, most traditional SEM fit indices to test for model fit are not available when using the LMS method in M*plus*, except for Akaike's information criterion (AIC; Akaike, 1998) and Bayesian information criterion (BIC; Schwarz, 1978).

In this chapter, we introduce the product indicator (PI) method to model latent interactions in RFA models. The PI method has received a great deal of attention in the general context of modeling interactions among latent variables in SEM (Henseler & Chin, 2010; Lin et al., 2010; Little et al., 2006; Marsh et al., 2004), but has never been studied in light of assessing measurement invariance. First, we discuss the assessment of measurement invariance using RFA models, then we introduce the PI method, and finally we demonstrate how to evaluate measurement invariance using RFA with PI by means of an illustrative example. We compare the results of PI to LMS on the same simulated data set.

2.2 Background

2.2.1 Restricted Factor Analysis

In RFA models, the construct T can be modeled as a latent factor with multiple measures X (e.g., items) as observed indicators. The background variable V is added to the measurement model as an exogenous single-indicator latent variable and is allowed to covary with the common factor T. The background variable V may represent a grouping variable by using a dummy-coded indicator. The observed indicator scores X are modeled as

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\lambda} t_i + \boldsymbol{b} v_i + \boldsymbol{c} t_i v_i + \boldsymbol{\delta} \boldsymbol{\varepsilon}_i \tag{2.2}$$

where $\boldsymbol{x}_i$ is a vector of indicator scores, t_i is the common factor T score, v_i is a dummy code for group membership V, and $\boldsymbol{\varepsilon}_i$ is a vector of the residual scores of subject i. Moreover, the vector $\boldsymbol{\tau}$ contains intercepts, $\boldsymbol{\lambda}$ is a vector of factor loadings on the common factor T, and $\boldsymbol{\delta}$ is a vector of residual factor loadings. The vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ are of special interest and contain regression coefficients. A nonzero element in $\boldsymbol{b}$ or $\boldsymbol{c}$ indicates uniform or nonuniform DIF, respectively.

Figure 2.1 illustrates an example of an RFA model to assess measurement invariance using two anchor indicators (i.e., Indicator 1 and 2). The background variable V is modeled as a latent variable with a single indicator Y representing group membership. For visual simplicity, the measurement model of $T \times V$ is excluded from Figure 2.1, but those details are discussed in the following subsection. Measurement invariance can be examined by comparing the fit of an unconstrained model with several constrained models. In the unconstrained model, all indicators are regressed on V and $T \times V$, except for the

---

[2]LMS is also available in the open-source R package `nlsem` (Umbach et al., 2017), but the implementation is very limited. It is not possible to assess measurement invariance using RFA models in the `nlsem` package, so we do not consider it further.

Figure 2.1: An example of an RFA model with LMS for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to test Indicator 3 for uniform and nonuniform DIF, respectively.

indicators in the anchor set. Each constrained model involves fixing the regression of the studied indicator onto V and $T \times V$ at zero.

The pair of constraints for each indicator can be tested simultaneously, where the null hypothesis of measurement invariance implies both $\boldsymbol{b}$ and $\boldsymbol{c}$ coefficients corresponding to the studied indicator are zero in the population. These constraints can be tested via model comparison of a constrained and unconstrained model, producing a likelihood ratio test (LRT) statistic that is distributed as χ^2 random variable with $df = 2$. A significant LRT statistic indicates that the studied indicator exhibits DIF with respect to V, and 1-df follow-up tests of the individual b and c coefficients can reveal whether that indicator's DIF is uniform or nonuniform. This chapter focuses only on the 2-df omnibus test for each indicator.

2.2.2   Product Indicators

The use of PI to model interactions among latent variables was originated by Kenny & Judd (1984). The PI method involves the specification of a measurement model for the latent interaction factor. Generally, product terms are built by multiplying the indicators of the associated latent variables, which serve as indicators for the latent interaction factor. All indicators, including the product indicators, are assumed to be multivariate normally distributed if the maximum likelihood estimation procedure is used. Because products of normal variables are not themselves normally distributed, this assumption is violated. Thus, a robust maximum likelihood estimator is used to relax this assumption (see Marsh et al., 2004).

Several variants of the PI method have been proposed, among which is the double-mean-centering strategy (Lin et al., 2010) that we implement herein. The double-mean-centering strategy is superior to other strategies because it eliminates the need for a mean structure and does not involve a cumbersome estimation procedure. Although the orthogonalizing and double-mean-centering strategy perform equally well when all indicators are normally distributed, the double-mean-centering strategy performs better when the assumption that all indicators are normally distributed is violated (Lin et al., 2010).

The Double-Mean-Centering Strategy

The first step of the double-mean-centering strategy involves mean-centering the indicators of the latent variables of interest. Each of the mean-centered indicators of one latent variable is then multiplied by the mean-centered indicators of the other latent variable. The resulting product indicators are centered at their means and are used as indicators of the latent interaction factor. If the common factor T has P indicators and the background variable V has Q indicators, then the latent interaction factor can have up to $P \times Q$ product indicators, although matching schemes have been proposed to reduce the number of product indicators (Marsh et al., 2004). In RFA, however, these matching schemes would be irrelevant when the common factor only interacts with a single-indicator background variable (or with multiple single-indicator background variables). Figure 2.2 shows an example of an RFA model with a latent interaction using the PI method. All possible cross-products are used in this example (i.e., each indicator of T is multiplied by the single indicator of V), and all indicators of T and V are centered at their means[3].

2.3 Tutorial

To demonstrate how the PI method can be applied for assessing measurement invariance, we simulated a single data set. R syntax for the use of PI in RFA models is provided in the following subsection. Barendse et al. (2012) provide M*plus* syntax to implement LMS.

2.3.1 Data Generation

Data were generated for two groups, each with a group size of $n = 100$. We considered a scale of $P = 10$ indicators, 40% of which functioned differently: two indicators exhibited uniform DIF and two indicators exhibited nonuniform DIF. This way, we are able to investigate the performance of LMS and PI using a hypothetical scale with a substantial

---

[3]In the case of a dummy-coded indicator, the mean is the proportion of the sample in Group 1. Mean-centering does not affect the variance, so a 1-unit increase in a mean-centered dummy code still represents a comparison of Group 1 to Group 0, just as the original dummy code does.
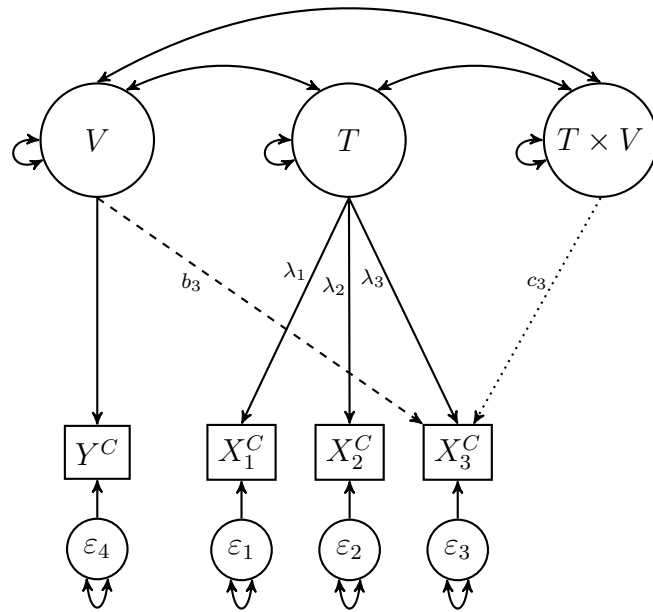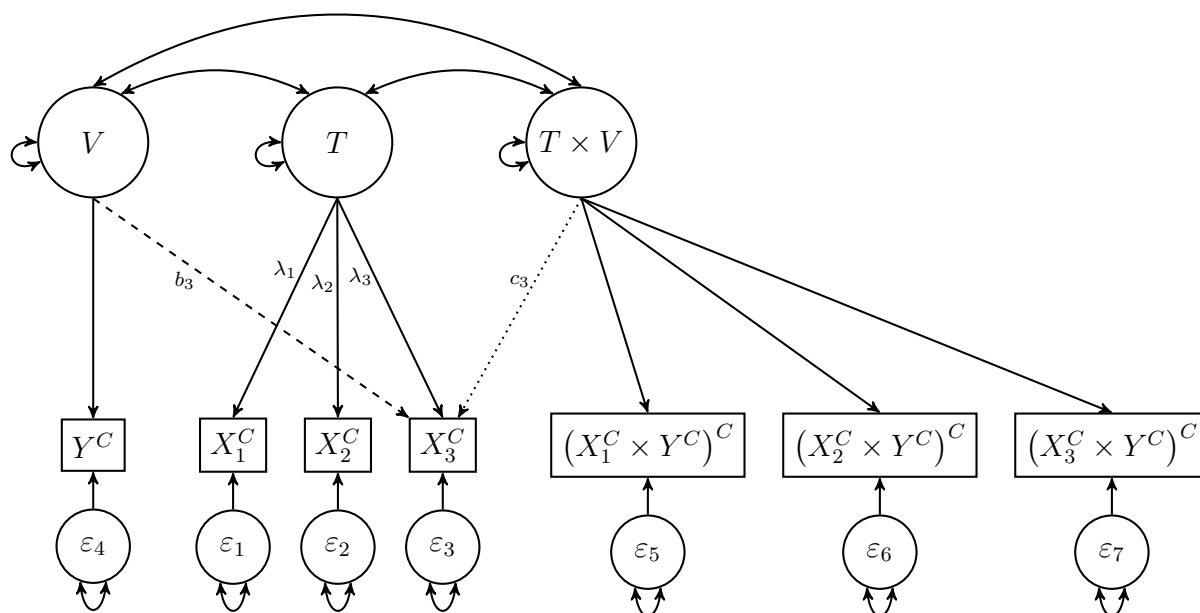
Figure 2.2: An example of an RFA model with product indicators for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to test Indicator 3 for uniform and nonuniform DIF, respectively.

degree of DIF. Indicator scores of subject i in group y were generated using the following model:

$$\boldsymbol{x}_i = \boldsymbol{\tau}_y + \boldsymbol{\lambda}_y t_i + \boldsymbol{\delta}_y \boldsymbol{\varepsilon}_i \tag{2.3}$$

where $\boldsymbol{x}_i$ is a vector of 10 indicator scores, t_i is the common factor score, and $\boldsymbol{\varepsilon}_i$ is a vector of 10 unique factor scores (residuals) for subject i. Moreover, $\boldsymbol{\tau}_y$ is a vector containing 10 intercepts, $\boldsymbol{\lambda}_y$ is a vector of 10 common factor loadings, and $\boldsymbol{\delta}_y$ is a vector of 10 residual factor loadings of group y. Following Barendse et al. (2010), differences in the common factor were simulated by drawing common factor scores t_i from a standard normal distribution $\mathcal{N}(0,1)$ for the reference group and from a normal distribution with a lower mean $\mathcal{N}(-0.5,1)$ for the focal group. Residual factor scores $\boldsymbol{\varepsilon}_i$ were drawn from a standard normal distribution.

The same magnitude of uniform and nonuniform DIF used by Barendse et al. (2010) was used. To introduce uniform DIF, all intercepts $\boldsymbol{\tau}$ were equal to 0, except for the intercept for the second and third indicator in the focal group, which were chosen equal to 0.5 (small uniform DIF) and 0.8 (large uniform DIF), respectively. Moreover, all common-factor loadings were fixed at 0.8, except for the factor loadings of the fourth and fifth indicator in the focal group, which were chosen equal to 0.55 (small nonuniform DIF) and 0.3 (large nonuniform DIF), respectively. The residual factor loadings were set equal to the square root of $1 - \boldsymbol{\lambda}_y^2$. Below we present the R syntax to generate this data set.

```
> ## set seed
> RNGkind("L'Ecuyer-CMRG")
> .Random.seed <- as.integer(c(407, 1945764513, -1852313839, 178524778,
> -983224279,-1572978333, -68534343))
```

```
> ## specify group size
> n <- 100
> ## draw latent-trait values
> trait1 <- rnorm(n)
> trait2 <- rnorm(n, -0.5, 1)
> ## draw scores on residual factor
> residual <- matrix(NA, 2*n, 10)
> for (j in 1:n) {
>    for (i in 1:10) {
>       residual[j, i] <- rnorm(1)
>    }
> }
> ## model parameters reference group
> lambda1 <- rep(0.8, 10)
> delta1 <- sqrt(1 - loading1^2)
> ## model parameters focal group
> tau2 <- c(0, -0.5, -0.8, 0, 0, 0, 0, 0, 0, 0)
> lambda2 <- c(0.8, 0.8, 0.8, 0.55, 0.3, 0.8, 0.8, 0.8, 0.8, 0.8)
> delta2 <- sqrt(1 - loading2^2)
> ## simulate indicator scores reference group
> x1 <- matrix(NA, n, 10)
> for (j in 1:n) {
>    for (i in 1:10) {
>       x1[j,i] <- lambda1[i] * trait1[j] + delta1[i] * residual[j, i]
>    }
> }
> ## simulate indicator scores focal group
> x2 <- matrix(NA, n, 10)
> for (j in 1:n) {
>    for (i in 1:10) {
>       x2[j,i] <- tau2[i] + lambda2[i]*trait2[j] + delta2[i]*residual[j,i]
>    }
> }
> ## combine scores of both groups
> dat <- as.data.frame(rbind(x1, x2))
> dat$group <- rep(c(1, 2), each = n)
> names(dat) <- paste0("x", 1:11)
```

2.3.2   Application

Below is the R syntax for the application of PI in RFA models to assess measurement invariance in the simulated data set. The RFA models with PI are fitted with the R package `lavaan` (version 0.5-23; Rosseel, 2012). In our example, we apply the double-mean-centering strategy. First, the `indProd()` function in the `semTools` package (version 0.4-14; Jorgensen et al., 2019) with the argument `doubleMC = TRUE` is used to transform the data in order to be suitable for this strategy. This way, the indicators of the common factor T and background variable V are mean-centered and indicators of the interaction factor $T \times V$ are built by multiplying the mean-centered indicator of V by each mean-centered indicator of T. The resulting product indicators are mean-centered again. After the data are prepared, one constrained model for each studied indicator must be specified. We use the ninth and tenth indicators, which are both DIF-free, as anchor indicators, so they are not tested for DIF. Hence, the studied indicators are the first eight indicators, four of which exhibit DIF, which leads to eight constrained models in total. The unconstrained model is the same across indicators.

```
> ## required package
> library(semTools)
> ## prepare data
> datDMC <- indProd(dat, 1:10, 11, match = FALSE, doubleMC = TRUE)
> ## additional parameters
> paramc <- paste0("group + group.by.trait =~ x", 1:8)
> ## specify and fit unconstrained model
> mod.un <- c('
>   theta =~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10
>   group =~ 1*x11
>   group.by.trait =~ x1.x11 + x2.x11 + x3.x11 + x4.x11 + x5.x11 +
>                     x6.x11 + x7.x11 + x8.x11 + x9.x11 + x10.x11
>   x11 ~~ 0*x11', paramc)
> mod.un.fit <- cfa(mod.un, data = datDMC, estimator = "MLM")
> ## specify and fit constrained models
> out <- matrix(NA, nrow = 8, ncol = 2,
>               dimnames = list(paste0("x", 1:8), c("X2", "p")))
> for (i in 1:length(paramc)) {
>   mod.con <- mod.un[-(i+1)] # remove b and c for the i-th studied indicator
>   mod.con.fit <- cfa(mod.con, data = datDMC, estimator = "MLM")
>   outfit <- lavTestLRT(mod.con.fit, mod.un.fit,
>                     method = "satorra.bentler.2001")
>   out[i,1:2] <- c(outfit[2,5], outfit[2,7])
> }
> ## print results
> out
```

The first factor of the unconstrained model is the common factor T with 10 mean-centered observed variables X^C as indicators. The second factor is the background variable V with a mean-centered single indicator Y^C representing group membership. The residual variance of Y^C is fixed at zero. The interaction factor $T \times V$ is the third factor of the unconstrained model with double-mean-centered product indicators. For example, the first indicator of the interaction factor is obtained by mean-centering $Y^C \times X_1^C$. For all factors in the unconstrained model, the factor loading λ of the first indicator is fixed at unity for identification. Covariances between all three factors are freely estimated. Finally, factor loadings of all indicators on V and $T \times V$ are added, except for the anchor indicators. The constrained models are built by removing factor loadings of the studied indicator on V and $T \times V$ from the unconstrained model. The estimator to be used for the unconstrained and constrained models is set to `"MLM"`, which involves maximum likelihood estimation with robust standard errors and a Satorra-Bentler scaled test statistic (Rosseel, 2012).

To test each of the eight indicators for DIF, likelihood ratio test statistics are calculated using the `lavTestLRT()` function in the `lavaan` package (version 0.5-23; Rosseel, 2012). This involves comparing the fit of the unconstrained model with each constrained model. By setting the argument `method = "satorra.bentler.2001"`, a scaled $\Delta\chi^2$ test statistic with $df = 2$ is computed as described by Satorra & Bentler (2001). An indicator is flagged as exhibiting DIF with respect to background variable V when the $\Delta\chi^2$ statistic is significant using a criterion of $\alpha = .05$.

Table 2.1 presents the results of measurement invariance assessment using RFA with LMS and PI. When the PI method was applied, the $\Delta\chi^2$ statistics of three out of four

truly DIF indicators were significant. The indicator with small nonuniform DIF, Indicator 4, was not flagged as exhibiting DIF, which is consistent with previous Monte Carlo studies showing that power to detect uniform DIF is greater than to detect nonuniform DIF (Barendse et al., 2010, 2012). Moreover, none of the $\Delta\chi^2$ statistics of the DIF-free indicators were significant. Thus, none of the indicators were incorrectly flagged using PI. The LMS method obtained comparable results, but correctly flagged all truly DIF indicators.

Table 2.1: Results of assessing measurement invariance using RFA models with PI and LMS.

Indicator	PI		LMS	
	$\chi^2_{df=2}$	p	$\chi^2_{df=2}$	p
1	0.425	.809	0.674	.714
2	**19.396**	**.000**	**17.696**	**.000**
3	**38.755**	**.000**	**28.000**	**.000**
4	5.212	.074	**6.283**	**.043**
5	**10.105**	**.006**	**10.656**	**.005**
6	0.145	.930	0.201	.904
7	0.948	.622	0.772	.680
8	0.246	.884	0.196	.907

Note. Bold cells indicate significant DIF. Indicators 9 and 10 were used as anchor indicators, so they were not tested for DIF.

2.4 Discussion

In this chapter, we proposed the use of PI in RFA models as an alternative to LMS to assess measurement invariance. The illustrative example showed that this method obtains results comparable to LMS. Because RFA with LMS can only be implemented in M*plus* (L. K. Muthén & Muthén, 2012), knowing that PI performs at least as well as LMS provides more researchers the opportunity to test for nonuniform DIF using any SEM software package. An additional advantage of PI is the availability of more traditional SEM fit indices to test for model fit that are not available when using LMS in M*plus*, nor when using other available strategies for modeling interactions with latent variables (e.g., random effects models which treat indicator scores as cross-nested within indicators and subjects). However, several aspects of the use of PI in RFA models are yet unclear, for example, which indicators should serve as product indicators for the interaction factor (e.g., all indicators, only anchor indicators, or anchor indicators and studied indicators). In addition, RFA models assume strict invariance, that is, equal residual variances across groups. Future research could investigate how violations of strict invariance affect Type I error rates.

Chapter 3

Using Restricted Factor Analysis to Select Anchor Indicators and Assess Measurement Invariance

Abstract

Restricted factor analysis (RFA) is a powerful method to assess scalar invariance, but it may require empirically selecting anchor indicators to prevent inflated Type I error rates. We conducted a simulation study to compare two empirical anchor-selection strategies: a one-step rank-based strategy and an iterative selection procedure. Unlike the iterative procedure, the rank-based strategy had a low risk and degree of contamination within the empirically selected anchor set, even with small samples. To detect violations of scalar invariance, RFA requires an interaction effect with the latent factor. The latent moderated structural equations (LMS) method has been applied to RFA and has revealed inflated Type I error rates. We propose using product indicators (PI) as a more widely available alternative to measure the latent interaction. A simulation study, involving several sample-size conditions and magnitudes of measurement invariance violations, revealed that PI obtained similar power but lower Type I error rates, as compared to LMS.

3.1 Introduction

In the absence of measurement invariance, observed differences in composite scores (e.g., scale means) might not represent true differences in the construct that a scale is developed to measure. Measurement invariance is formally defined (Mellenbergh, 1989):

$$f_1(X|T = t, V = v) = f_2(X|T = t) \qquad (3.1)$$

where X is a set of observed variables measuring the construct of interest T, and V is a set of variables other than T that potentially violate measurement invariance (e.g., groups defined by sex or ethnicity). Throughout this chapter, we will use the term indicator to refer to the observed indicators X of the construct T, and refer to variable V as background variable. Function $f_1(\cdot)$ is the conditional distribution of X given T and V, and $f_2(\cdot)$ the conditional distribution of X given T. If measurement invariance holds (i.e., $f_1 = f_2$), the measurement of T by X is invariant with respect to V. If measurement invariance does not hold (i.e., $f_1 \neq f_2$), however, the measurement of T by X functions differently with respect to V. A violation of measurement invariance is commonly referred to as differential item functioning (DIF). A distinction can be made between violations of scalar invariance and metric invariance, also called uniform and nonuniform DIF, respectively. Uniform DIF implies that the magnitude of DIF is constant for all levels of the construct T, whereas nonuniform DIF implies that the magnitude of DIF varies with T. In different measurement contexts, DIF goes by many other names, such as measurement bias (Oort, 1992), noninvariance (Byrne et al., 1989), or differential indicator functioning (Kline, 2011, p. 253).

A common method to assess measurement invariance with respect to a grouping variable V is multiple-group confirmatory factor analysis (MGCFA; Vandenberg & Lance, 2000), in which a measurement model is estimated for each group, and then invariance constraints are imposed on the parameter estimates in order to test whether any indicators exhibit DIF. Hence, this method requires sufficiently large samples for each group. Restricted factor analysis (RFA; Oort, 1992, 1998) is an alternative when sample sizes are small. In RFA models, the background variable V is added to a measurement model as an exogenous variable that is allowed to covary with T. Multiple-indicator multiple-cause (MIMIC) models (B. O. Muthén, 1989) are statistically equivalent to RFA models, but instead of a covariance between V and T, a causal effect of V on T is modeled. An advantage of RFA over MGCFA is that the division of the sample into subsamples by V is not necessary, but RFA also involves an additional assumption—namely, homogeneity of common and unique factor variances across groups. If this additional assumption hold, RFA has slightly higher power than MGCFA to detect DIF (Barendse et al., 2012).

A possible disadvantage of RFA is that it is not readily suited to assess metric invariance. Because a violation of metric invariance implies that the magnitude of DIF varies as a function of the common factor T, an interaction effect of T with V on X should be estimated. To this end, RFA has been extended with a distribution-analytic approach to

model interactions in factor models called latent moderated structural equations (LMS; Barendse et al., 2010). With LMS the background variable V should be modeled as a single-indicator latent variable in the RFA model (or MIMIC model; Woods & Grimm, 2011) to enable estimation of a latent interaction of T with V, thus allowing nonuniform DIF to be estimated as the latent-interaction's effect(s) on the indicator(s). Barendse et al. (2010, 2012) showed that RFA with LMS generally has high power (89% to 100%) to detect both uniform and nonuniform DIF, except in conditions with a small sample size and small nonuniform DIF. However, severely inflated Type I error rates have been observed (Barendse et al., 2010, 2012; Woods & Grimm, 2011). This motivated us to find an alternative method for estimating the interaction effect of T with V on X that would provide better control of Type I error rates.

The first aim of this chapter was to compare the performance of LMS with that of product indicators (PI; Kenny & Judd, 1984), which is an alternative method to model interactions between latent variables in structural equation models. We aimed to examine whether this method can minimize the inflated Type I error rates obtained with LMS when assessing measurement invariance using RFA models. The PI method has been studied extensively in the general context of modeling latent interactions in structural equation modeling (SEM; Henseler & Chin, 2010; Lin et al., 2010; Little et al., 2006; Marsh et al., 2004), but its performance in RFA models to assess metric invariance has not yet been explored. An advantage of PI over LMS is that it can be implemented in any SEM software package, and several methods for calculating product indicators have been automated in the open-source R package `semTools` (version 0.5-0; Jorgensen et al., 2018). In contrast, assessing metric invariance with RFA models using LMS can only be applied with the commercial SEM software M*plus* (L. K. Muthén & Muthén, 2012). In addition to its limited availability, this software does not provide most traditional SEM fit indices to test for model fit when using LMS estimation. A preliminary study on the use of PI in RFA models suggests that PI and LMS obtain comparable conclusions about whether an indicator exhibits (non)uniform DIF (Kolbe & Jorgensen, 2018). However, a more extensive simulation study was necessary to (dis)confirm the promising performance of the PI method in RFA models for assessing measurement invariance.

Methods for assessing measurement invariance generally require the selection of anchor indicators. These indicators are indicators used to link the scales of the latent construct of interest across groups and are assumed to be DIF-free. A common strategy is to use all indicators other than the studied indicator as anchors. This strategy leads to a contaminated subset of anchor indicators when some indicators other than the studied indicator exhibit DIF, which in turn leads to problems such as inaccurate indicator-parameter estimates and an overestimation of the amount of DIF in the test data (W.-C. Wang, 2004). Hence, Woods (2009) argued that the inflated Type I error rates obtained with LMS might be caused by a contaminated subset of anchor indicators. A simulation study of Woods & Grimm (2011) showed that LMS still resulted in inflated Type I error rates when using an uncontaminated anchor set, which calls into question whether any alternative method

might control Type I errors better, given a valid set of anchor indicators.

The importance of an uncontaminated anchor set for assessing measurement invariance provided a second motivation for this chapter: to investigate practical methods of empirically identifying anchor indicators when they are not known a priori. Rather than explicitly selecting anchor indicators, Barendse et al. (2012) applied RFA with LMS to assess measurement invariance, iteratively accounting for violations of measurement in one indicator at a time. They showed that this brings Type I error rates closer to the nominal level of significance, although some inflation remains. In the present study, we adapted the iterative procedure suggested by Barendse et al. (2012) as an anchor-selection strategy, to be implemented before testing for DIF—that is, iteratively removing indicators from an anchor set initially consisting of all indicators. The iterative procedure can arguably result in large anchor sets, because it begins by assuming all indicators as anchors and then selects indicators to remove from this anchor set. The potential danger of a larger anchor set is that it generally displays a higher risk of contamination than a smaller anchor set (Kopf et al., 2015b). Therefore, we contrasted the iterative procedure with the rank-based strategy proposed by Woods (2009). This is an easily implemented forward-selection strategy, in which a limited proportion of all indicators—those that show the weakest evidence against measurement invariance—are added to the anchor set. A similar strategy has already been applied in MIMIC models (Chun et al., 2016). Woods (2009) recommended that the number of indicators in the anchor set should be approximately 10% to 20% of the total number of indicators.

We will describe both our adaptation of Barendse et al.'s (2012) iterative procedure and Woods' (2009) rank-based strategy for empirically selecting anchor indicators in greater detail in a later section. Because both of these empirical anchor-selection strategies involve preliminary assessments of measurement invariance, we begin by describing how to assess measurement invariance using RFA models with both LMS and PI. A description of anchor-selection strategies follows, after which we describe two simulation studies: one to compare anchor-selection strategies and the other to compare latent-interaction models for assessing measurement invariance.

3.2   Background

3.2.1   Restricted Factor Analysis

The data-generating model for observed continuous scores x with potential uniform and nonuniform DIF can be written as follows

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\lambda} t_i + \boldsymbol{b} v_i + \boldsymbol{c} t_i v_i + \boldsymbol{\delta} \boldsymbol{\varepsilon}_i \tag{3.2}$$

where $\boldsymbol{x}_i$ is a vector of observed scores, t_i is the common factor T score, v_i is the background variable V score, and $\boldsymbol{\varepsilon}_i$ is a vector of the residual scores of subject i. The

background variable V can be either observed or latent, continuous or categorical[1], and it is allowed to covary with the common factor T. The model parameters in Equation 3.2 include a vector of intercepts $\boldsymbol{\tau}$, a vector of factor loadings $\boldsymbol{\lambda}$ on the common factor T, a vector of residual factor loadings $\boldsymbol{\delta}$, and vectors of regression coefficients $\boldsymbol{b}$ and $\boldsymbol{c}$. The regression coefficients in $\boldsymbol{b}$ represent the linear effect of the background variable V on the observed scores $\boldsymbol{x}_i$, and a nonzero element in $\boldsymbol{b}$ indicates uniform DIF (i.e., a violation of scalar invariance). The regression coefficients in $\boldsymbol{c}$ represent the nonlinear interaction effects of V with T on $\boldsymbol{x}_i$, and a nonzero element in $\boldsymbol{c}$ indicates nonuniform DIF (i.e., a violation of metric invariance).

When an MGCFA model is fitted to sample data generated under the population described by Equation 3.2, $\boldsymbol{b}$ and $\boldsymbol{c}$ are not explicitly estimated, but their effects are implicitly captured by virtue of allowing $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$, respectively, to vary across levels of V. In contrast, a single-group RFA model for the common factor T with observed indicators X can be fitted to the data, where the background variable V is added to the model as an exogenous variable. The analytical RFA model resembles the data-generating Equation 3.2, but it fixes $\boldsymbol{\delta} = 1$ for identification. Furthermore, traditional maximum likelihood estimation of an RFA model is complicated by the inability to calculate the product between an observed background variable V and the latent T in order to estimate the nonlinear interaction effects $\boldsymbol{c}$. LMS has been proposed as a solution to model these nonlinear interaction effects in RFA models (e.g., Barendse et al., 2010), and we have proposed PI as a more widely available alternative method (Kolbe & Jorgensen, 2018), which we investigated more thoroughly in the current study.

In general, scalar and metric invariance can be assessed through RFA by comparing the fit of an unconstrained model with the fit of a constrained model. The unconstrained model freely estimates b and c parameters for all indicators studied (i.e., nonanchor indicators), fixing the b and c parameters of the anchor indicators at zero. In the constrained model, the b and c parameters of a single studied indicator are additionally fixed at zero. Any potential DIF in other to-be-studied indicators is controlled for, because the b and c parameters of those indicators are freely estimated in both models. This minimizes the chance of inflated Type I error rates (Woods & Grimm, 2011).

For each studied indicator, the constraints on the b and c parameters can be tested simultaneously via model comparison of that indicator's constrained model with the unconstrained model. This comparison produces a likelihood ratio test (LRT) statistic, which is distributed as a χ^2 random variable with $df = 2$. A significant LRT statistic indicates that the studied indicator functions differently with respect to V. To reveal whether this DIF is uniform or nonuniform, follow-up tests of the individual b and c coefficients can be performed using each parameter's Wald z statistic. We focused our investigation only on the omnibus test with $df = 2$ for each studied indicator.

---

[1]Equation 3.2 could be expanded with additional dummy effects or contrast codes when V has > 2 categories. Additional background variables could also be added to Equation 3.2 to reflect additive or interactive effects on the measurements.

Latent Moderated Structural Equations

RFA has most commonly been extended with LMS to assess metric invariance (Barendse et al., 2010, 2012; Woods & Grimm, 2011). The LMS approach to estimate interaction effects of latent variables is a distributional analytic approach available in M*plus* (L. K. Muthén & Muthén, 2012), which implements a maximum likelihood estimation procedure developed especially for the distributional properties of a model that includes product terms among normally distributed latent factors (A. Klein & Moosbrugger, 2000). In LMS, the joint distribution of indicators is represented as a finite mixture of normal distributions. The mixture distribution function is used in order to obtain maximum likelihood estimates by means of the expectation maximization algorithm (Dempster et al., 1977). The LMS approach assumes multivariate normality for all latent exogenous variables. The most common situation for testing invariance is a comparison of two groups (Putnick & Bornstein, 2016), but when the background variable V is a categorical variable, the normality assumption is violated. This violation can be accounted for by using a robust maximum likelihood estimator (Woods & Grimm, 2011). Additional details on how to apply LMS in M*plus* and an example M*plus* script for fitting the RFA model with LMS are provided by Barendse et al. (2012).

Figure 3.1 depicts an example RFA model estimable with LMS for assessing measurement invariance. The model represents a ten-indicator case, with the last two indicators treated as anchors. The LMS approach requires the background variable V to be modeled as a latent variable. In this example, the background variable is measured by a single indicator Y representing group membership. As indicated in Figure 3.1, the residual variance of Y has to be fixed at zero in order for the model to be identified. LMS uses the raw data of all indicators in the model for estimation, but it does not require any indicators of the latent interaction factor $T \times V$; hence, that factor is represented by a dotted circle. Measurement invariance can be assessed for each indicator by comparing the fit of an unconstrained model with the fit of a constrained model. The unconstrained model regresses all studied indicators on V and $T \times V$, but not anchor indicators (in Figure 3.1, the indicators X_9 and X_{10} are anchors, not regressed on V and $T \times V$). Put differently, the b and c parameters only of the anchor indicators are fixed at zero. In a constrained model, the b and c parameters of a studied indicator are additionally set to zero, to assess measurement invariance for that indicator.

Product Indicators

Another possibility for estimating the nonlinear interaction effects in RFA models is the PI method proposed by Kenny & Judd (1984). The PI method involves specifying a measurement model for an additional factor, referred to as the latent interaction factor, which represents the interaction between two latent variables. Hence, using the PI method in RFA models requires the background variable V to be modeled as a latent variable, and the measurement model of the latent interaction factor is specified using products between

Figure 3.1: An RFA model with LMS for assessing measurement invariance. The indicators X_9 and X_{10} are the anchor indicators. The dashed and dotted arrows represent effects that may be estimated in order to assess scalar and metric invariance, respectively.

the background variable and each of the indicators of T. If maximum likelihood is used to estimate the parameters of a model with product indicators, all indicators (including the product indicators) are assumed to be multivariate normally distributed. This assumption is violated because even products of normal variables are not normally distributed; the present example, however, involves the product of normal indicators with a binary dummy code, which is itself not normally distributed. A robust maximum likelihood estimator should therefore be used (Marsh et al., 2004).

There are various PI methods that differ in the formation of the product indicators of the latent interaction factor. The most recently proposed PI method is the double-mean-centering strategy (Lin et al., 2010). Using this strategy, indicators for the latent interaction factor are built by mean-centering the product terms produced by multiplying the mean-centered indicators of the associated latent variables. In our ten-indicator example with a grouping variable as the background variable, an initial product term between the grouping variable Y and an indicator (e.g., the first indicator X_1) is first calculated from the mean-centered variables[2]: $(Y - \bar{Y})(X_1 - \bar{X}_1)$. The double mean-

---

[2] The "mean" of a binary dummy code is the proportion of the sample for whom the dummy code equals 1, so the mean-centered dummy code will still have only 2 levels: the zeros become negative and the ones become positive. This transformation does not change the interpretation of its effect because the distance between the negative value and positive value is still one unit. That is, its effect on an indicator is interpreted as the average change in that indicator associated with a one-unit change in the (mean-centered) dummy code, which therefore still represents the difference between two groups' means (controlling for other predictors, such as T). Because the double-mean-centering strategy negates the need for modeling a mean structure, mean-centering a dummy code does not affect the interpretation of any (covariance-structure) model parameters.
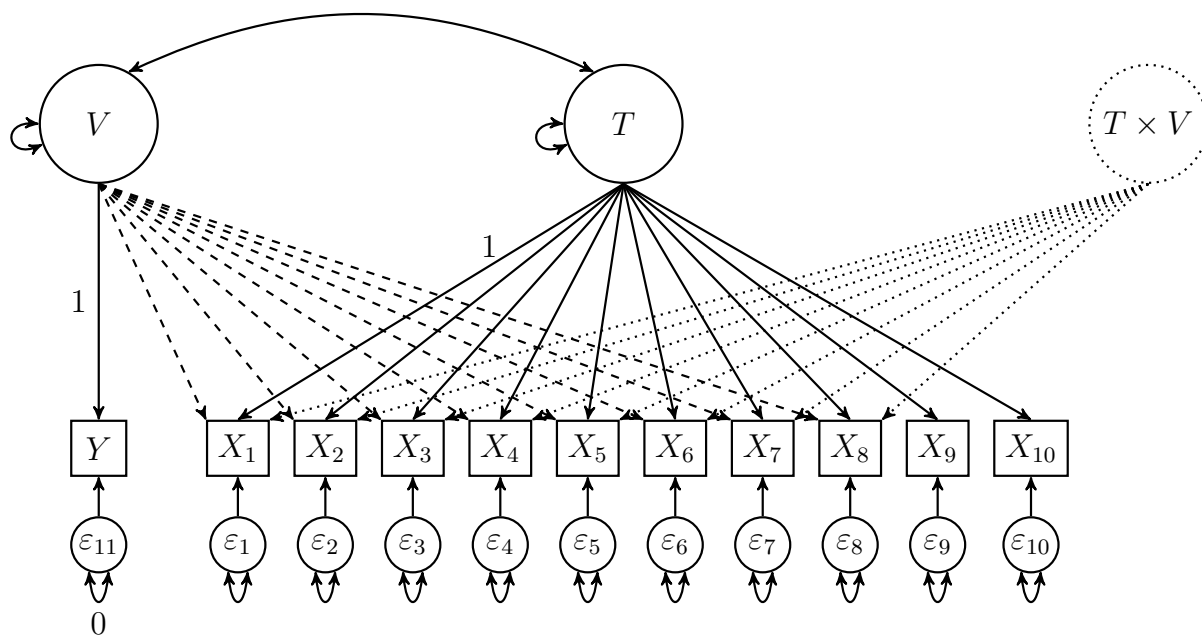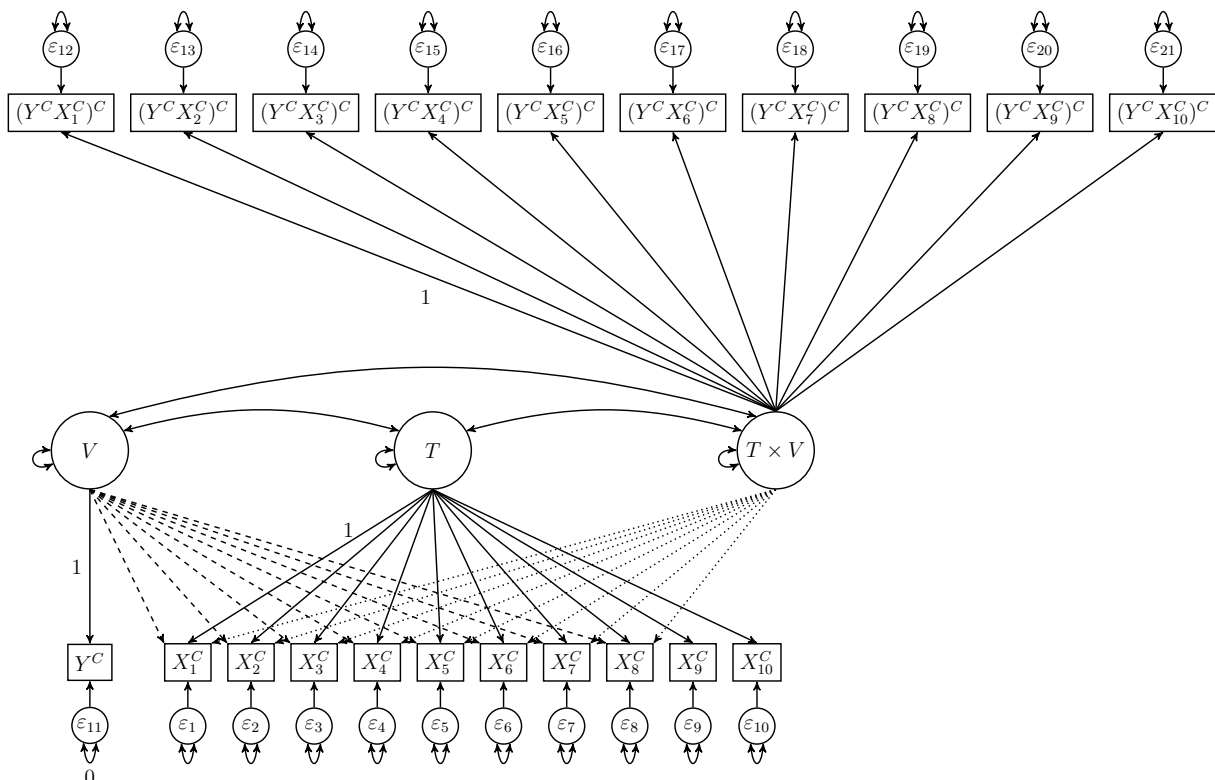
Figure 3.2: An RFA model with product indicators for assessing measurement invariance. The indicators X_9 and X_{10} are the anchor indicators. The dashed and dotted arrows represent effects that may be estimated to assess scalar and metric invariance, respectively.

centered product indicator is then formed by mean-centering the initial product term: $(Y - \bar{Y})(X_1 - \bar{X}_1) - \overline{(Y - \bar{Y})(X_1 - \bar{X}_1)}$.

Advantages of the double-mean-centering strategy over other strategies are that it does not require a mean structure to be modeled and does not involve a cumbersome multistep estimation procedure. An additional advantage is that this strategy outperforms other strategies when the assumption of normality is violated (Lin et al., 2010). See Kolbe & Jorgensen (2018) for an example application of RFA using double-mean-centered product indicators in the R package `lavaan` (Rosseel, 2012).

Figure 3.2 illustrates the same ten-indicator example of an RFA model, but the latent interaction factor $T \times V$ is estimated with a measurement model using product indicators calculated via the double-mean-centering strategy. This example includes ten mean-centered indicators, of which the last two are treated as anchors. Each mean-centered indicator of T is multiplied by the mean-centered indicator of V, and all indicators of $T \times V$ are recentered in order to obtain the double mean-centered product indicators. The double-mean-centered product indicators are denoted as $(Y^C X_p^C)^C$ for the $p = 1, \ldots, 10$ indicators in Figure 3.2.

Although the path diagram in Figure 3.2 still represents the statistical model fitted to the data, it should not be interpreted as representing an actual data-generating model. The $T \times V$ factor is not an independently identified latent variable, nor are its indicators,

so their factor loadings should not be interpreted as the causal effects of $T \times V$ on the product indicators. Product indicators are calculated from other variables in the model, and their loadings merely represent the portion of a product indicator's variance associated with the product of T with V, as opposed to the product of V with that indicator's unique factor. Thus, the "measurement" of a latent $T \times V$ factor is merely an ad hoc technique for extracting the variance that is in common among all of the (double-mean-centered) indicators, so that the effects of the latent $T \times V$ factor on actual indicators (i.e., indicators of T) can be estimated in order to assess metric invariance.

As in the LMS method, measurement invariance can be assessed for each indicator by comparing the fit of an unconstrained model (i.e., the b and c parameters of only the anchor indicators are fixed at zero) with the fit of a constrained model (i.e., the b and c parameters of the studied indicator are additionally set to zero). Unlike the ad hoc interpretation of the factor loadings for $T \times V$, the interpretation of c_p is straightforward: the degree to which the effect of latent factor T on observed variable X_p is moderated by V.

3.2.2 Anchor-Selection Strategies

An anchor-selection strategy guides the decision about which particular indicators should be used as anchors when assessing measurement invariance on indicator-level. The anchor indicators are presumed to be DIF-free and are used to identify the latent construct (i.e., the model would not be identified if all indicators loaded on T and were regressed on V and $T \times V$, as well as estimating factor covariances). In RFA models, anchor indicators are not regressed on V and $T \times V$ when assessing studied indicators for measurement invariance. Several strategies for selecting anchor indicators have been proposed. Some strategies rely on prior knowledge of DIF-free indicators or content experts' advice, whereas empirical strategies are based on preliminary indicator analysis. This study focused only on empirical anchor-selection strategies. We first describe Woods's (2009) rank-based strategy because it involves fewer steps than Barendse et al.'s (2012) iterative procedure.

Rank-Based Strategy

The rank-based strategy introduced by Woods (2009) involves an easy procedure to select anchor indicators. It stems from the idea that the value of each indicator's test statistic reflects the magnitude of DIF of that indicator. The proposed strategy is to assess all indicators for measurement invariance using all other indicators as anchors. A test statistic with $df = 2$ can be calculated to assess measurement invariance for one indicator at a time. In the context of RFA, the fit of a constrained model can be compared with the fit of several unconstrained models (one per indicator). In the constrained model, none of the indicators is regressed on V or $T \times V$, whereas in each unconstrained model, only the studied indicator is regressed on V and $T \times V$. After calculating a test statistic for each indicator's set of constraints, the indicators are ranked in ascending order based on

their test statistics. The indicators with the smallest test statistics (i.e., weakest evidence against measurement invariance) are selected as anchor indicators. The actual number of indicators being selected as anchor indicators may be determined by factors such as test length and sample size. Woods (2009) suggested that the number of indicators selected as anchors should be approximately 10-20% of the total number of indicators.

Iterative Procedure

The iterative procedure was proposed by Barendse et al. (2012) as a method to detect violations of measurement invariance, but can also be applied to select anchor indicators (for examples, see Candell & Drasgow, 1988; Hidalgo-Montesinos & Lopez-Pina, 2002; Kopf et al., 2015a,b). Similar to the rank-based strategy, this procedure involves comparing the fit of a constrained model with several unconstrained models. In the constrained model, none of the indicators is regressed on V and $T \times V$, whereas in an unconstrained model, a studied indicator is regressed on V and $T \times V$. Instead of choosing anchors among the indicators with the weakest evidence against measurement invariance, all indicators are initially considered eligible as anchors, and the indicators with the strongest evidence against measurement invariance are removed from consideration. In the first run of the iterative procedure, the indicator with the largest significant test statistic is considered to function differently. This DIF is taken into account in the second iteration by allowing for the regression of that indicator on V and $T \times V$ in the constrained and unconstrained models. The remaining indicators are assessed for measurement invariance, and again, the indicator with the strongest significant evidence against measurement invariance is removed from consideration. The constrained and unconstrained models are again modified by regressing this indicator on V and $T \times V$ before testing the remaining indicators. This procedure continues until none of the remaining indicators has a significant test statistic, or until half of the indicators are considered to function differently. Any remaining indicators are considered DIF-free and used as anchor indicators when assessing all other indicators (again[3]) for measurement invariance.

3.3   Study 1: Selecting Anchor Indicators

3.3.1   Method

In this study, we used simulated data to examine the suitability of the rank-based strategy (Woods, 2009) and the iterative procedure (Barendse et al., 2012) to select anchor indicators. The suitability of these strategies was assessed in the context of extending

---

[3]Recall that measurement invariance is assessed by comparing one constrained model per indicator to the same unconstrained model (which constrains b and c only for anchor indicators). This is distinct from the approach used by anchor-selection strategies, which compare one unconstrained model per indicator to the same constrained model (which constrains b and c for all indicators). The former approach is preferred for the assessment of measurement invariance to prevent inflating Type I error rates (Woods & Grimm, 2011).

RFA with both LMS and PI. In addition to the latent-interaction method (LMS vs. PI), we manipulated anchor-selection strategy (rank-based with 20% or 70% as anchors, or iterative procedure), group sample size ($n = 50, 100, 150$, or 200 per group), and size of DIF (small or large), yielding a $2 \times 3 \times 4 \times 2$ factorial design with 1000 replications in each condition. The relatively small group sample sizes were used because that is the situation when RFA is preferred over MGCFA, which requires larger samples (Oort, 1998). Our outcomes included risk of contamination (i.e., the percentage of replications in which the anchor set contained at least one indicator exhibiting DIF) and degree of contamination (i.e., the percentage of selected anchor indicators within each set that exhibited DIF), for which we report the means in each condition.

Data Generation

Data were generated for two groups under different sample sizes. A scale of $P = 10$ indicators was considered, of which one indicator exhibited uniform DIF, one indicator exhibited nonuniform DIF, and one indicator exhibited both types of DIF. This allowed us to investigate the performance of the anchor-selection strategies under nonideal conditions because a substantial degree of contamination in the anchor set was possible. The following model was used to generate indicator scores of subject i in group y:

$$\boldsymbol{x}_i = \boldsymbol{\tau}_y + \boldsymbol{\lambda}_y t_i + \boldsymbol{\delta}_y \boldsymbol{\varepsilon}_i \qquad (3.3)$$

where $\boldsymbol{x}_i$ is a vector of 10 indicator scores, t_i is subject i's common factor score, and $\boldsymbol{\varepsilon}_i$ is a vector of residual factor scores of subject i. Differences in common factor scores between the groups were simulated by drawing common factor scores t_i from a standard normal distribution $\mathcal{N}(0, 1)$ for the reference group and from a normal distribution with a lower mean and variance $\mathcal{N}(-0.5, 0.7)$ for the focal group, similar to (Barendse et al., 2010). Residual factor scores $\boldsymbol{\varepsilon}_i$ in both groups were drawn from a standard normal distribution $\mathcal{N}(0, 1)$.

The group-specific vector $\boldsymbol{\tau}_y$ includes 10 intercepts, and $\boldsymbol{\lambda}_y$ includes 10 common factor loadings. We replicated the same magnitude of uniform and nonuniform DIF used by (Barendse et al., 2010). Uniform DIF was introduced by imposing across-group differences in intercepts. All intercepts were equal to 0, except for the intercept for the second and fourth indicator in the focal group, which were equal to 0.5 in the small DIF-size conditions and 0.8 in the large DIF-size conditions. All common factor loadings were equal to 0.8, except for the factor loadings of the third and fourth indicator in the focal group, which were equal to either 0.55 or 0.3 in the conditions with small and large DIF, respectively. For each group g, the vector of residual factor loadings $\boldsymbol{\delta}_y$ was set equal to $\sqrt{1 - \boldsymbol{\lambda}_y^2}$ so the indicators had population variances equal to 1.

Analytical Procedure

Using RFA with both LMS and PI, measurement invariance was assessed for each indicator by comparing the fit of a constrained model with the fit of an unconstrained model. In the constrained model, b and c (see Equation 3.2) are vectors containing zeros, whereas in the unconstrained model, the elements in b and c corresponding to the studied indicator are freely estimated. The difference in fit between the models was compared with a robust χ^2 statistic with $df = 2$ (Satorra & Bentler, 2010), using $\alpha = .05$ as criterion for significance. In order to enable the estimation of the model parameters, group membership was modeled as a latent factor with a single indicator whose factor loading was fixed at unity in each of the models. The residual variance of the group membership indicator was fixed at zero in the RFA models with PI (see Figure 3.2), whereas this residual variance was fixed at 0.001 in the RFA models with LMS to overcome identification problems. For both methods, the common factor T was identified by fixing the factor loading of the first indicator at unity. In RFA models with PI, the factor loading of the first indicator of the interaction factor $T \times V$ was also fixed at unity.

When the iterative procedure was used select anchor indicators, indicators were iteratively assessed for measurement invariance. After each iteration, the indicator associated with the largest significant χ^2 test statistic was considered to function differently, and this DIF was explicitly modeled in the following iteration. The procedure continued until none of the remaining indicators was associated with a significant χ^2 statistic, or until half (i.e., five) of the indicators were considered to function differently. Any remaining indicators considered DIF-free after the final iteration were selected for the anchor set. If the χ^2 statistic of one or more of the studied indicators could not be determined (for instance, because of convergence problems), the procedure was ended and indicators considered DIF-free in the previous iteration were selected as anchor indicators. With these criteria, the iterative procedure could select 50–100% of the total number of indicators as anchors.

With the rank-based strategy, all indicators were assessed for measurement invariance and ranked in ascending order based on their χ^2 statistics. Then, the indicators with the lowest χ^2 statistic were selected for the anchor set. We examined two versions of the rank-based strategy, one in which 20% of the total number of indicators was selected as anchor indicator, as suggested by Woods (2009). In order to assess the effect of using a larger anchor set, and to fairly compare the results of the rank-based strategy to the iterative procedure by having a larger anchor set, we also examined the rank-based strategy when selecting seven indicators (70% of the total number of indicators) with the lowest statistic.

In each condition, the risk of contamination was determined, which represents the percentage of replications that yielded an anchor set containing at least one indicator with DIF. In addition to the risk of contamination, we evaluated the degree of contamination in the anchor set, which is the percentage of indicators exhibiting DIF in the anchor set. For both risk and degree of contamination, we report the mean across replications in each condition. Because the iterative procedure may result in varying lengths of anchor sets,

Table 3.1: Percentage of replications using LMS with invalid results in Study 1

	Percentage of invalid results	
n	Small DIF	Large DIF
50	23.10	22,50
100	18.00	15.50
150	19.50	17.40
200	24.40	26.30

Note. The total number of replications in each condition was 1000. Only the percentages of invalid results when using LMS were reported in this table, because none of the replications with product indicators obtained invalid results.

the average count (i.e., the average number of indicators with DIF in the anchor set) of this selection strategy was calculated. The RFA models with LMS were fit with M*plus* (version 7; L. K. Muthén & Muthén, 2012) via the R package M*plus*Automation (version 0.7; Hallquist & Wiley, 2018). The RFA models with PI were fit with the R package `lavaan` (version 0.5-23; Rosseel, 2012). Results were analyzed with R (version 3.3.2; R Core Team, 2016).

3.3.2 Results

After conducting the analysis for each of the conditions, we found that the LMS method did not always produce valid results due to convergence problems. The percentages of replications with invalid results in each condition with the LMS method are represented in Table 3.1. Across all conditions, convergence problems occurred in 20.84% of all replications using LMS. The convergence problems did not seem to be associated with either the sample size of the groups or the size of DIF. Among the cases with convergence problems, one or more indicators could not be assessed for measurement invariance because the χ^2 statistic(s) for the corresponding indicator(s) could not be calculated. When such problems occurred with the LMS method using the rank-based strategy, the results of that replication in that condition were not included in the analysis, because an unambiguous decision about anchor indicators could not be made in practice. Similarly, when convergence problems occurred with the LMS method in the first run of the iterative procedure, the replication in that condition was not included in the analysis.

In contrast, each of the models converged for every single replication among all conditions using PI. Because the results of the LMS method were based on a smaller number of replications, the validity of comparing results between the two methods could be considered questionable (e.g., if the subsample of replications for which LMS had convergence problems was not a random sample from all 1000 replications, at least with respect to our outcomes of interest). Therefore, we also calculated results for the PI method using

Table 3.2: Results of the anchor-selection strategies for each of the conditions in Study 1

Method	Size of DIF	n	Average risk of contamination			Average degree of contamination		
			RB(20%)	RB (70%)	IP	RB(20%)	RB (70%)	IP
LMS	Small	50	6.11	62.68	88.30	3.06 (0.061)	9.23 (0.646)	14.14 (1.203)
		100	1.46	36.83	73.05	0.73 (0.015)	52.26 (0.368)	11.39 (0.973)
		150	0.62	23.35	55.78	0.31 (0.006)	3.34 (0.234)	9.43 (0.826)
		200	0.13	14.68	43.52	0.07 (0.001)	2.10 (0.147)	8.14 (0.735)
	Large	50	0.52	24.52	52.77	0.26 (0.005)	3.52 (0.247)	8.77 (0.764)
		100	0.12	4.26	26.04	0.06 (0.001)	0.61 (0.043)	5.91 (0.556)
		150	0.00	0.85	17.68	0.00 (0.000)	0.12 (0.008)	4.64 (0.452)
		200	0.95	0.95	20.90	0.47 (0.009)	0.14 (0.009)	5.93 (0.585)
PI	Small	50	5.80	65.70	89.50	2.90 (0.058)	9.66 (0.676)	12.33 (0.995)
		100	2.00	43.40	75.30	1.00 (0.020)	6.20 (0.434)	9.53 (0.753)
		150	0.40	31.80	57.30	0.20 (0.004)	4.54 (0.318)	7.29 (0.573)
		200	0.30	21.30	43.70	0.15 (0.003)	3.04 (0.213)	5.54 (0.437)
	Large	50	2.90	31.80	49.60	1.45 (0.029)	4.63 (0.324)	6.38 (0.506)
		100	0.20	7.30	11.00	0.10 (0.002)	1.04 (0.073)	1.42 (0.111)
		150	0.10	2.20	0.50	0.05 (0.001)	0.31 (0.022)	0.07 (0.005)
		200	0.00	0.40	0.20	0.00 (0.000)	0.06 (0.004)	0.02 (0.002)

Note. The average count (i.e., the average number of DIF indicators in the anchor set) is reported in parentheses alongside the average degree (percentage) of contamination. LMS = latent moderated structural equations; PI = product indicators; RB (20%) = rank-based strategy selecting 20% of all indicators as anchors; RB (70%) = rank-based strategy selecting 70% of all indicators as anchors; IP = iterative procedure. Risk of contamination = percentage of replications in which the anchor set contained at least one indicator exhibiting DIF; Degree of contamination = percentage of indicators exhibiting DIF in the anchor set averaged over all replications.

only the replications for which LMS converged. We found the same pattern of results when comparing methods using only the replications that had no convergence problems, so below we present results using all available converged replications in each condition.

Risk of Contamination

Table 3.2 shows the risk and degree of contamination of the selection strategies within each condition. Across all conditions, the rank-based strategy selecting 20% of the total number of indicators as anchors had the lowest risk of contamination compared to the rank-based strategy selecting 70% of the indicators as anchors and the iterative procedure. The rank-based strategy selecting 20% of the total number of indicators as anchors had a risk of contamination of 0.00% to 6.11%, whereas the rank-based strategy selecting 70% of the indicators as anchors had a risk of contamination ranging from 0.40% to 65.70%. The selection strategy with the highest risk of contamination in each of the conditions was the iterative procedure, except in conditions using PI where the size of DIF was large and the sample size was either 150 or 200. Among all conditions, the iterative procedure had a risk of contamination of 0.20% to 89.50%. The risk of contamination generally decreased with sample size and size of DIF for each selection strategy. For example, the risk of contamination for the iterative procedure with small DIF and $n = 200$ was less than half of the risk of contamination with small DIF when $n = 50$.

Table 3.3: Average number of indicators selected as anchors in the iterative procedure

		Number of indicators in the anchor set	
Size of DIF	n	LMS	PI
Small	50	8.112	7.897
	100	7.878	7.671
	150	7.739	7.476
	200	7.631	7.350
Large	50	7.679	7.415
	100	7.460	7.034
	150	7.366	6.929
	200	7.479	6.920

Note. LMS = latent moderated structural equations; PI = product indicators. Ideally, only seven indicators would be included in the anchor set (i.e., the seven indicators without DIF in the population).

Degree of Contamination

Similar to the risk of contamination, the degree of contamination typically decreased with sample size and size of DIF for each selection strategy. The rank-based strategy selecting 20% of the indicators as anchors had the lowest degree of contamination in the majority of the conditions, with an overall degree of contamination of 0.68%. The only condition in which the rank-based strategy selecting 70% of the indicators as anchors had a lower degree of contamination was when using LMS with a large size of DIF and a sample size of $n = 200$. In this condition, the rank-based strategy selecting 20% of the indicators as anchors yielded a degree of contamination of 0.47%, whereas the rank-based strategy selecting 70% as anchor indicators had a degree of contamination of 0.14%. In all other conditions, the rank-based strategy selecting 20% of the indicators as anchors performed better than the two other selection strategies with respect to the degree of contamination. In addition, the iterative procedure had the highest degree of contamination in the majority of the conditions. To ensure a fair comparison between the anchor-selection strategies on the outcome variables, the average number of anchor indicators selected by the iterative procedure for each sample size and size of DIF condition is reported in Table 3.3. The average number of indicators selected as anchors by the iterative procedure was typically close to seven (ranging from 6.920 to 8.112 across conditions), which was comparable to the number of indicators selected by the rank-based strategy in the 70% condition.

3.4 Study 2: Assessing Measurement Invariance

3.4.1 Method

In Study 2, we evaluated the Type I error rates and power of the LMS and PI methods in RFA models to detect violations of scalar and metric invariance. In addition to the

latent-interaction method (LMS vs. PI), we again manipulated the reference and focal group sample sizes ($n = 50$, 100, 150, or 200 per group) and the size of DIF (small or large), but not anchor-selection strategy. Because Study 1 showed that the rank-based strategy selecting 20% of the total number of indicators as anchors yielded the lowest risk and degree of contamination in the anchor set, only two out of 10 indicators were used as anchors in Study 2. But we manipulated an additional factor (known vs. unknown anchors). The performance of LMS and PI was assessed in the best-case scenario; that is, two known DIF-free indicators (Indicators 9 and 10) were used as anchor indicators and were not assessed for measurement invariance. This best-case scenario always yielded a DIF-free anchor set. By comparison, we also used an empirical-selection scenario, in which the two anchor indicators selected by the rank-based strategy from Study 1 were used as anchors to assess all other indicators for measurement invariance. This yielded a $2 \times 3 \times 4 \times 2$ factorial design, using the same random-number seeds to generate the same 1000 data sets in each sample-size and DIF-size condition of Study 1.

Analytical Procedure

Measurement invariance was assessed for each indicator by comparing the fit of an unconstrained model with the fit of several constrained models (one per studied indicator) using a robust χ^2 statistic with $df = 2$ (Satorra & Bentler, 2010). In the unconstrained model, all elements in $\boldsymbol{b}$ and $\boldsymbol{c}$ were freely estimated, except for the elements corresponding to the anchor indicators. For all studied indicators, a constrained model was fitted, in which the corresponding elements in $\boldsymbol{b}$ and $\boldsymbol{c}$ for the studied indicator were fixed at zero. The same identification constraints were used as in Study 1. An indicator was flagged as an indicator with DIF when the χ^2 statistic was significant at $\alpha = .05$.

Power and Type I error rates were calculated across all conditions. Power reflects the proportion of replications in which the truly DIF indicators were correctly flagged as indicator with DIF. The Type I error rate represents the proportion of replications in which there was at least one Type I error (i.e., one of the DIF-free indicators was incorrectly flagged as indicator with DIF). Agresti-Coull confidence intervals [4] (Agresti & Coull, 1998) around the observed Type I error rates were calculated to evaluate the significance of inflation. Power was calculated for each type (uniform, nonuniform, and both) and magnitude (small and large) of DIF separately. The models were fit with M*plus* (version 7; L. K. Muthén & Muthén, 2012) via the `MplusAutomation` package (version 0.7; Hallquist & Wiley, 2018) in the LMS conditions and `lavaan` (version 0.5-23; Rosseel, 2012) in the PI conditions, and results were analyzed with R (version 3.3.2; R Core Team, 2016).

---

[4]Agresti-Coull confidence intervals were obtained by first defining $\tilde{J} = J + z^2$, where J is the total number of replications in a single condition and z the $1 - \alpha/2$ quantile of a standard normal distribution. Then, the midpoint for E Type I errors is determined by $\tilde{p} = \frac{1}{\tilde{J}}(E + \frac{z^2}{2})$. The Agresti-Coull confidence interval around the Type I error rate is given by $\tilde{p} \pm \sqrt{\frac{\tilde{p}}{\tilde{J}}(1 - \tilde{p})}$.

Table 3.4: Percentage of replications with invalid results in Study 2 for the best-case and empirical scenarios

			Percentage of invalid results	
Method	Size of DIF	n	Best-case	Empirical
LMS	Small	50	21.50	24.40
		100	18.10	19.70
		150	26.50	22.50
		200	31.40	26.40
	Large	50	21.50	24.10
		100	18.00	18.30
		150	25.60	22.60
		200	32.50	28.90

Note. The total number of replications in each condition was 1000. Only the percentages of invalid results when using LMS were reported in this table, because none of the replications with product indicators obtained invalid results.

3.4.2 Results

After performing the analysis for each of the conditions, we again observed a number of replications with invalid results when using the LMS method. Table 3.4 shows the percentages of replications with invalid results among the conditions for the best-case scenario and empirical scenario. On average across all best-case scenario conditions, invalid results were obtained in 24.39% of all replications using LMS. For each of these replications, the problem involved a non-converging unconstrained model. Due to this complication, a χ^2 statistic could not be calculated for any of the indicators. The results of these replications in the best-case scenario were not included in the analysis because in practice, a researcher would not be able to assess measurement invariance in this situation using RFA. The empirical scenario obtained invalid results in 23.36% of all replications averaged across the conditions with LMS. These replications were excluded from the analysis for this scenario because in practice, a decision could not be made regarding the selection of anchor indicators, or measurement invariance could not be assessed due to a non-converging unconstrained model.

The PI method did not produce any convergence problems. All models converged for every replication in each condition. Because the analysis of the LMS method included a smaller number of replications, we again compared results between the two methods using only the replications for which LMS converged. The same pattern of results was found for this smaller set of replications, so we present results using all available converged replications in each condition.

Table 3.5: Power of the LMS and PI method under each condition of the best-case scenario in Study 2

Type of DIF	n	Small DIF		Large DIF	
		LMS	PI	LMS	PI
Uniform	50	.828	.737	.932	.981
	100	.960	.995	.977	1.000
	150	.973	1.000	.991	1.000
	200	.994	1.000	.991	1.000
Nonuniform	50	.162	.108	.535	.464
	100	.341	.218	.834	.839
	150	.544	.358	.882	.973
	200	.672	.493	.825	.996
Combination	50	.660	.710	.925	.977
	100	.947	.994	.966	1.000
	150	.980	1.000	.974	1.000
	200	.993	1.000	.982	1.000

Note. LMS = latent moderated structural equations; PI = product indicators; small uniform DIF = a difference of 0.5 in intercepts across groups; large uniform DIF = a difference of 0.8 in intercepts across groups; small nonuniform DIF = a difference of 0.25 in factor loadings across groups; large nonuniform DIF = a difference of 0.5 in factor loadings across groups.

Best-Case Scenario

Table 3.5 shows the power of LMS and PI across conditions in the best-case scenario (always a DIF-free anchor set). In the majority of the conditions, the PI method obtained a higher power than LMS, although the differences were quite small. Exceptions included the power to detect small nonuniform DIF, which was higher for LMS than for PI. In contrast, large nonuniform DIF was more often detected by PI than by LMS. Power generally increased with sample size for all types and sizes of DIF. Relative to uniform DIF, nonuniform DIF was more difficult to detect, which is consistent with previous research (Barendse et al., 2010). Both LMS and PI especially yielded low power for small nonuniform DIF. With a sample size of $n = 50$, for example, small nonuniform DIF was only detected in 10.80% to 16.20% of all replications. Moreover, the power to detect indicators exhibiting both uniform and nonuniform DIF was in most conditions comparable to the power for uniform DIF.

Type I error rates for the LMS method in the best-case scenario ranged between .080 and .200 (see, Table 3.6). In each of the conditions, the error rates were significantly larger than the nominal level of significance (5%). The Agresti-Coull confidence intervals for the error rates in each condition with LMS were above the nominal level of significance. The PI method yielded Type I error rates ranging from .047 to .068. When $n = 100$, the error rates were above the nominal level of significance, and the Agresti-Coull lower confidence limits for these error rates were just above the nominal level of significance. In

Table 3.6: Type I error rates of LMS and PI under each condition of
the best-case scenario in Study 2

Size of DIF	n	Type I error [95% CI]	
		LMS	PI
Small	50	**.088 [.070, .110]**	.058 [.045, .074]
	100	**.129 [.108, .154]**	**.068 [.054, .085]**
	150	**.151 [.127, .179]**	.051 [.039, .067]
	200	**.197 [.169, .228]**	.047 [.035, .062]
Large	50	**.080 [.063, .101]**	.059 [.046, .075]
	100	**.133 [.111, .158]**	**.069 [.055, .087]**
	150	**.149 [.125, .177]**	.051 [.039, .067]
	200	**.200 [.167, .227]**	.049 [.037, .064]

Note. LMS = latent moderated structural equations; PI = product in-
dicators. **Bold** font indicates the lower 95% confidence limit exceeds the
nominal 5% alpha level, implying the Type I error rate is statistically sig-
nificantly inflated. The square brackets contain Agresti-Coull confidence
intervals around the error rates.

the remaining conditions, error rates were slightly above or below $\alpha = .05$. However, the
error rates for these conditions were not significantly smaller or larger than .05 because
Agresti-Coull confidence intervals for the Type I error rates in conditions with $n = 50$,
$n = 150$, and $n = 200$ included the nominal level of significance. Because there is no
reason to expect only the $n = 100$ condition to yield (barely) inflated error rates, we
assume this only reflects Monte Carlo sampling error.

Empirical Scenario

Table 3.7 shows the power of LMS and PI across conditions in the empirical scenario in
which two anchor indicators were selected with the rank-based strategy. The pattern of
results found for the empirical scenario was comparable to the best-case scenario. For
example, similar to the best-case scenario, the PI method had more power to detect DIF
than LMS in the majority of the conditions, but the differences were generally small.
Again, a noticeable exception was the power to detect small nonuniform DIF, which was
higher for LMS than for PI. With a sample size of $n = 50$, small nonuniform DIF was
only detected by PI in 5.70% of all replications.

Type I error rates for the LMS method in the empirical scenario ranged from .077 to
.247 (see Table 3.8). As in the best-case scenario, each of the error rates of LMS was
significantly larger than the nominal level of significance. The Agresti-Coull confidence
intervals for these error rates were entirely above the nominal level of significance. By
comparison, the Type I error rates for the PI method ranged from .022 to .048. In the
condition with large DIF and a sample size of $n = 50$ or $n = 100$, the Agresti-Coull
confidence interval around the error rate included the nominal level of significance. The
confidence intervals of the other conditions were all below $\alpha = .05$.

Table 3.7: Power of the LMS and PI method under each condition of the empirical scenario in Study 2

		Small DIF		Large DIF	
Type of DIF	n	LMS	PI	LMS	PI
Uniform	50	.718	.560	.906	.949
	100	.963	.979	.969	.998
	150	.983	1.000	.994	.999
	200	.997	1.000	.992	1.000
Nonuniform	50	.168	.057	.573	.422
	100	.367	.156	.859	.825
	150	.563	.288	.894	.980
	200	.696	.414	.834	.997
Combination	50	.577	.537	.920	.948
	100	.949	.977	.963	.999
	150	.974	.999	.984	1.000
	200	.997	1.000	.997	1.000

Note. LMS = latent moderated structural equations; PI = product indicators; small uniform DIF = a difference of 0.5 in intercepts across groups; large uniform DIF = a difference of 0.8 in intercepts across groups; small nonuniform DIF = a difference of 0.25 in factor loadings across groups; large nonuniform DIF = a difference of 0.5 in factor loadings across groups.

Table 3.8: Type I error rates of LMS and PI under each condition of the empirical scenario in Study 2

		Type I error [95% CI]	
Size of DIF	n	LMS	PI
Small	50	**.077 [.060, .098]**	.022 [.014, .033]
	100	**.105 [.085, .128]**	.026 [.018, .038]
	150	**.129 [.107, .155]**	.023 [.015, .034]
	200	**.247 [.217, .280]**	.035 [.025, .048]
Large	50	**.083 [.065, .105]**	.048 [.036, .063]
	100	**.106 [.087, .130]**	.041 [.030, .055]
	150	**.123 [.101, .148]**	.032 [.023, .045]
	200	**.231 [.201, .263]**	.032 [.023, .045]

Note. LMS = latent moderated structural equations; PI = product indicators. **Bold** font indicates the lower 95% confidence limit exceeds the nominal 5% alpha level, implying the Type I error rate is statistically significantly inflated. The square brackets contain Agresti-Coull confidence intervals around the error rates.

3.5 Discussion

The present study concerned assessing measurement invariance using RFA models. One of the aims of this study was to compare LMS with PI, an alternative method to model latent interactions. We examined whether this method can minimize the inflated Type I error rates obtained with LMS when assessing measurement invariance using RFA models. Woods (2009) argued that the inflated Type I error rates of LMS might be caused by a contaminated set of anchor indicators. Hence, prior to the comparison between the two methods to model latent interactions, we investigated which anchor-selection strategy is most suitable when assessing measurement invariance using RFA models.

The findings of Study 1 indicate that Wood's (2009) rank-based strategy selecting a small number of indicators as anchors is more suitable than an iterative procedure of removing indicators with DIF from the anchor set (Barendse et al., 2012). The rank-based strategy selecting 20% of the total number of indicators as anchors consistently yielded lower risk and degree of contamination and performed well across all sample sizes. These results are in line with previous studies (M. Wang & Woods, 2017; Woods, 2009), which showed that the rank-based strategy frequently obtains a DIF-free anchor set. The most striking finding of Study 1 is perhaps the high risk of contamination yielded by the rank-based strategy when selecting 70% of the total number of anchor indicators and by the iterative procedure. These selection strategies allow for larger anchor sets, which generally display a higher risk of contamination than smaller anchor sets (Kopf et al., 2015b). It is also worth noting that other promising empirical anchor-selection strategies have been identified in the IRT literature that could also generalize well to RFA (or multigroup CFA)—namely, the forward mean test-statistic threshold and forward mean p-value threshold methods (Kopf et al., 2015a)—but their implementation is not as straight-forward as the rank-based strategy, which yielded excellent results even with small samples. Future research could focus on identifying optimal anchor-selection strategies for factor analysis models in various contexts (e.g., MGCFA).

In Study 2, we compared the LMS and PI methods to model latent interactions in RFA models. The main conclusion is that PI obtained similar power but lower Type I error rates compared to LMS. In line with previous studies, severely inflated Type I error rates were observed in conditions with LMS (Barendse et al., 2010, 2012; Woods & Grimm, 2011). Although it has been argued that the inflated Type I error rates obtained with LMS might be caused by a contaminated set of anchor indicators (Woods, 2009), our results contradict this possible explanation. The severely inflated error rates were not only observed in the empirical scenario in which contamination of the anchor set was allowed, but also in the best-case scenario with a DIF-free anchor set. This suggests that a contaminated anchor set may not fully account for the frequently observed inflated error rates when using LMS. In response to a reviewer's suggestion to increase the external validity of our Monte Carlo design, we allowed factor variances to differ across groups. This could explain why our Type I error rates under LMS were larger than those reported

by Barendse et al. (2010, 2012) and Woods & Grimm (2011), given further support by Chun et al. (2016) recent demonstration that unequal factor variances yield more inflated Type I error rates than equal factor variances when using LMS.

In contrast, the Type I error rates observed in conditions with PI were all close to the nominal level of significance in the best-case scenario of a DIF-free anchor set, and slightly below the nominal level of significance when using empirically selected anchors. Hence, the results of the current study indicate that the PI method can minimize the inflated Type I error rates obtained with LMS. We suspect a possible explanation for PI's better control of errors could be the explicitly estimated covariance between the latent factor T and interaction $T \times V$, which is not a free parameter in LMS estimation algorithms (A. Klein & Moosbrugger, 2000). This warrants further investigation, but is beyond the scope of the current investigation.

Corresponding to findings of previous studies (Barendse et al., 2010, 2012), we found that nonuniform DIF was more difficult to detect than uniform DIF. Power to detect nonuniform DIF was especially low in conditions with a small sample size. This finding is concerning to some extent, because the present investigation included a best-case scenario of a DIF-free anchor set. As opposed to simulation studies where the indicators with true DIF are known, in practice there may seldom be any reliable prior knowledge about DIF in the indicators of a scale. The results of the empirical scenario, however, show that empirically selecting anchor indicators using the rank-based strategy selecting 20% of the total number of indicators has minor impact on the assessment of measurement invariance. The power to detect DIF using an empirically selected anchor set with this strategy was comparable to the power observed in the best-case scenario. A possible explanation for this minor impact is that the selection strategy used in the empirical scenario yielded a remarkably low risk and degree of contamination in Study 1. Future research could more extensively investigate the consequences of different anchor-selection strategies on power and Type I error in the context of RFA.

An additional limitation of the LMS method brought to light by the present study is the large proportion of invalid results due to convergence problems. These convergence problems point to an important practical limitation of the LMS method, because in practice, a researcher would be unable to make a decision about anchor indicators or to assess indicators for measurement invariance. Moreover, this study showed that the PI method to model latent interactions in RFA models generally performs at least as well as the LMS method for the purpose of assessing measurement invariance. Because RFA extended with LMS can only be applied in the commercial SEM software M*plus* (L. K. Muthén & Muthén, 2012), knowing that PI is a viable alternative to LMS provides more researchers with the opportunity to assess metric invariance using RFA with any SEM software package. However, several aspects of the use of PI are yet unclear, for example, which indicators should serve as product indicators for the interaction factor. There are various possibilities regarding the formation of product indicators, among others are using only the studied indicator, only the anchor indicators, the anchor indicators

and the studied indicator, or all indicators (the latter of which was employed in the current study). Although this study showed promising results, more research is necessary to determine the optimal use of PI in RFA models for assessing measurement invariance.

Chapter 4

The Impact of Unmodeled Heteroskedasticity on Assessing Measurement Invariance

Abstract

This study compared two single-group approaches for assessing measurement invariance across an observed background variable: restricted factor analysis (RFA) and moderated nonlinear factor analysis (MNLFA). In MNLFA models, heteroskedasticity can be accounted for by allowing the common-factor variance and the residual variances to differ as a function of the background variable. In contrast, RFA models assume homoskedasticity of both the common factor and the residuals. We conducted a simulation study to examine the performance of RFA and MNLFA under common-factor and residual homoskedasticity and heteroskedasticity. Results suggest that MNLFA and RFA with product indicators outperform RFA with latent moderated structural equations in conditions with heteroskedastic common factors, and MNLFA outperforms RFA in conditions with residual heteroskedasticity. We provide an explanation for the robustness of RFA with product indicators to violations of common-factor homoskedasticity.

4.1 Introduction

Research in the social and behavioral sciences commonly depends upon measures of constructs that are not directly observable. In order to meaningfully compare measurements of latent constructs across individuals or groups, measurement invariance is required. Measurement invariance is formally defined as

$$f_1(X|T,V) = f_2(X|T), \tag{4.1}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ denote probability distributions, X is a set of observed variables (also referred to as indicators in this chapter) measuring the latent construct of interest T, and V is a set of background variables that are a potential source of a violation of measurement invariance (Mellenbergh, 1989). If measurement invariance holds, the measurement X depends only on the latent construct T and is invariant with respect to other variables V. However, if measurement invariance does not hold (i.e., $f_1 \neq f_2$), the measurement X depends not only on the latent construct T but also on V. With a lack of measurement invariance, individuals with an equal standing on the latent construct may have different expected values of X, and differences in the observed-score means may not represent true differences in T. Hence, before comparing measures of a latent construct, it is important to test the assumption of measurement invariance.

The majority of studies about measurement invariance involve omnibus tests for all of a particular type of measurement parameter (i.e., factor loadings or intercepts; see Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Finch & French, 2018; Marsh, 1994), as described below. But much less advice is available on how researchers should proceed when they reject an omnibus null hypothesis. Byrne et al. (1989) introduced the idea that partial invariance is sufficient to compare groups on their common-factor distributions. In the absence of a strong theory to specify a priori partial-invariance models to be tested, establishing partial invariance requires exploring which indicators' measurement parameters differ as a function of V. In some cases (e.g., many groups, no obvious reference group), recently proposed alignment (B. Muthén & Asparouhov, 2018; Marsh et al., 2018) or projection methods (Deng & Yuan, 2016; Jiang et al., 2017) may offer a promising way to compare latent distributions without explicitly locating violations of invariance. But when comparing very specifically chosen groups (e.g., men and women, clinical and healthy populations), it might be of great substantive interest to discover and explain why some indicators function differently across groups (or across a continuous V such as age), with important implications for how a scale or test is used in practice. When researchers have such interest, an analysis of indicator-level measurement invariance or differential item functioning (DIF)—as is more frequently discussed in the context of item-response theory (IRT) than structural equation modeling (SEM)[1]—could be indispensably informative.

A commonly used method to assess measurement invariance with respect to a cate-

---

[1]Exceptions include Suh (2015); Masyn (2017); Kolbe & Jorgensen (2019).

gorical variable V is multiple-group confirmatory factor analysis (MGCFA; Vandenberg & Lance, 2000). In MGCFA, a confirmatory factor model is simultaneously estimated for each group in which the construct T is modeled as a common factor with multiple indicators X, and invariance constraints are imposed on the parameter estimates in order to assess increasingly restrictive levels of measurement invariance (Meredith, 1993). Invariance can be tested for multiple factors without loss of generality, but we focus on the context of a single-factor model (Mellenbergh, 1994) to keep the discussion concise. The least restrictive level of invariance, called configural invariance, implies that the same factor structure holds across different levels of V. A more restrictive level of invariance is metric invariance, reflected by equality of the factor loadings across different levels of V. Yet more restrictive is scalar invariance, which posits that in addition to the factor loadings, each indicator's intercept is also equal across V. Additionally constraining residual variances (i.e., the variance of each indicator's unique factor) to equality across V is referred to as strict invariance.

An alternative method for evaluating measurement invariance with respect to a categorical variable V is restricted factor analysis (RFA; Oort, 1992, 1998). RFA models are single-group confirmatory factor models in which T is modeled as a common factor with multiple measures X as indicators, and V is included as an exogenous variable that freely covaries with T. To test whether scalar invariance is violated with respect to a particular X, X is regressed on V, and that slope represents a difference in intercepts of X across levels of V. RFA is thus readily suited to assess (violations of) scalar invariance, but assessing metric invariance requires estimating an interaction effect of T with V on X (i.e., different loadings across V implies that V moderates the effect of T on X). This interaction can be modeled in several ways, including the distribution-analytic approach called latent moderated structural equations (LMS; Barendse et al., 2010). Although RFA with LMS has high power to detect violations of scalar and metric invariance, several studies observed severely inflated Type I error rates (Barendse et al., 2010, 2012; Woods & Grimm, 2011). An alternative to LMS for estimating the interaction effect of T with V on X is the product indicator (PI; Kenny & Judd, 1984) method. Studies showed that the PI method generally performs well with respect to bias, precision, power, and Type I error rates in the context of modeling latent interactions in SEM (Henseler & Chin, 2010; Lin et al., 2010; Little et al., 2006; Marsh et al., 2004). Most recently, Kolbe & Jorgensen (2018) proposed the use of PI in RFA models to assess metric invariance. A simulation study on RFA with PI has shown that this method obtains similar power but more acceptable Type I error rates than LMS (Kolbe & Jorgensen, 2019).

There are several advantages of RFA over MGCFA. As the data are aggregated over subsamples in RFA models, RFA may provide higher power than MGCFA to detect violations of measurement invariance (Barendse et al., 2012). Another advantage of RFA over MGCFA is that it easily accommodates tests for measurement invariance with respect to a continuous variable V. In MGCFA models, testing for measurement invariance with respect to a continuous variable would require the continuous variable V to be catego-

rized, which can lead to a loss of power and measurement reliability (MacCallum et al., 2002). However, RFA comes with the additional assumptions of equal common-factor variances across different levels of V (i.e., common-factor homoskedasticity) and equal indicators' residual variances across different levels of V (i.e., residual homoskedasticity). The robustness of RFA to common-factor heteroskedasticity is relatively unexplored (see Chun et al., 2016; Harpole, 2015, for exceptions). Chun et al. (2016) studied the effect of common-factor heteroskedasticity with a categorical V on assessing measurement invariance using multiple-indicator multiple-cause (MIMIC) models, which are statistically equivalent to RFA models. Their study showed that Type I error rates were inflated as a result of common-factor heteroskedasticity. A more extensive study is required to examine whether the performance of RFA (or MIMIC) varies as a function of different magnitudes of factor-variance differences across V. The robustness of RFA to residual heteroskedasticity has also not yet been explored in depth, however, it has been argued that residual heteroskedasticity has similar impacts as common-factor heteroskedasticity (Meredith & Teresi, 2006).

When common-factor variances are suspected to differ with V, moderated nonlinear factor analysis (MNLFA) models may be a more suitable alternative to RFA for assessing measurement invariance. MNLFA was developed by Bauer & Hussong (2009, but see the earlier work by e.g., Neale, 1998; Neale et al., 2006; Mehta & Neale, 2005) and described as a tool for measurement invariance assessment by Bauer (2017). Similar to RFA, MNLFA does not require dividing the sample into subsamples by V, therefore also allowing for a continuous V. In MNLFA models, measurement invariance is examined in a single-group confirmatory factor model by means of parameter moderation. The variable V may alter the values of any subset of parameters including the common-factor variance and residual variances of the indicators X. As such, MNLFA does not require assuming common-factor or residual homoskedasticity with respect to V. The use of MNLFA for assessing measurement invariance has been evaluated with empirical data (see Bauer, 2017; Hildebrandt et al., 2016), and a simulation with categorical indicators showed that it performs well in large samples (e.g., $N = 2000$) when combined with a regularization approach (Bauer et al., 2020). However, its statistical properties (e.g., Type I error rates and power) have not yet been compared to other methods or investigated in simulation studies including conditions with small samples and continuous indicators.

The aim of the present study was to compare the Type I error rates and power of different single-group methods to test for measurement invariance with respect to a categorical or a continuous V. We conducted a Monte Carlo simulation study to evaluate the performance of RFA and MNLFA under common-factor and residual homoskedasticity and heteroskedasticity. The current study built on earlier work by Kolbe & Jorgensen (2019) for RFA models—as well as by Chun et al. (2016) for MIMIC models—but more extensively examined the impact of heteroskedasticity of both the common-factor and indicators' residuals on assessing metric and scalar invariance. That is, we investigated different magnitudes and directions of common-factor and residual variance differences,

and we simulated conditions with either a categorical or continuous variable V. Additionally, we contrasted not only LMS and PI within RFA models, but we also contrasted RFA with MNLFA models.

Following the results of previous studies (Chun et al., 2016; Kolbe & Jorgensen, 2019; Harpole, 2015), common-factor heteroskedasticity was hypothesized to inflate Type I errors using RFA with LMS to assess measurement invariance. We expected no impact of common-factor heteroskedasticity using RFA with PI because Kolbe & Jorgensen (2019) did not observe inflated Type I error rates despite common-factor variances being unequal. **Appendix I** offers an explanation for the robustness of the PI approach to violations of common-factor homoskedasticity. Although residual heteroskedasticity appears relatively unexplored in the context of RFA (or MIMIC), we held similar hypotheses about its inflation of Type I error rates, although we were unsure whether its impact would be as severe as that of common-factor heteroskedasticity.

The remainder of the chapter is organized as follows. First, we briefly describe RFA with LMS and PI, followed by a description of the MNFLA method for assessing measurement invariance. Then we present a Monte Carlo simulation study to compare these methods under various conditions. The chapter concludes with advice for applied researchers and suggestions for future research.

4.2   Background

We will start by considering the general form of a single-group confirmatory factor model. The basic principle of single-group models is that a set of common factors is modeled as being drawn from a single multivariate-normal distribution with a constant mean vector and covariance matrix for the entire population from which data were sampled. As mentioned above, we focus on a single-factor model. In a single-group model, the construct of interest T is operationalized as a latent factor with multiple observed measures X as indicators. Assuming continuous indicators X, the general form of a single-group model may be written as

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda} t_i + \boldsymbol{\varepsilon}_i, \tag{4.2}$$

where $\boldsymbol{x}_i$ is a $P \times 1$ vector of P observed indicator scores for person i, $\boldsymbol{\tau}$ is a $P \times 1$ vector of indicator intercepts, $\boldsymbol{\Lambda}$ is a $P \times 1$ vector of factor loadings, t_i is the common-factor score for person i and $\boldsymbol{\varepsilon}_i$ is a $P \times 1$ vector of residual scores for person i.

If measurement invariance holds with respect to a background variable V, the observed indicators X are affected directly only by the latent construct T, and only indirectly by V via T. Metric invariance requires equal $\boldsymbol{\Lambda}$ with respect to V, and scalar invariance additionally requires equal intercepts $\boldsymbol{\tau}$. In order to evaluate metric and scalar measurement invariance in a single-group model, the model for continuous indicators X can be

rewritten as

$$\begin{aligned}
\boldsymbol{x}_i &= \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i t_i + \boldsymbol{\varepsilon}_i \\
&= (\boldsymbol{\tau}_0 + \boldsymbol{b} v_i) + (\boldsymbol{\Lambda}_0 + \boldsymbol{c} v_i) t_i + \boldsymbol{\varepsilon}_i,
\end{aligned} \tag{4.3}$$

where v_i is the background variable score for person i, $\boldsymbol{\tau}_0$ is a $P \times 1$ vector of baseline intercepts when subject i's score on the variable V is $v_i = 0$, and $\boldsymbol{\Lambda}_0$ is a $P \times 1$ vector of baseline factor loading when $v_i = 0$. The $P \times 1$ vectors $\boldsymbol{b}$ and $\boldsymbol{c}$ are of special interest, because they contain coefficients that reflect violations of measurement invariance (i.e., DIF). A nonzero element in $\boldsymbol{b}$ implies a difference in an indicator's intercept τ with respect to V, and thus represents a violation of scalar invariance (called uniform DIF in the IRT literature). Similarly, a nonzero element in $\boldsymbol{c}$ implies an indicator's factor loading differs with respect to V, violating metric invariance (called nonuniform DIF).

The evaluation of scalar and metric invariance is thus concerned with testing the significance of the coefficients $\boldsymbol{b}$ and $\boldsymbol{c}$. For each indicator, an omnibus test of metric and scalar invariance can be conducted by comparing the fit of a constrained model with the fit of an unconstrained model. In the unconstrained model, all elements in $\boldsymbol{b}$ and $\boldsymbol{c}$ are freely estimated, except for the indicators that serve as anchors (i.e., indicators that are known or assumed to be invariant, rather than tested). In the constrained model for a particular tested indicator, that indicator's b and c are additionally fixed to zero, implying invariance of that indicator's measurement parameters. Any potential violation of measurement invariance in the other to-be-tested indicators is accounted for because the elements in $\boldsymbol{b}$ and $\boldsymbol{c}$ of those indicators are freely estimated in both models. The model comparison produces a likelihood ratio test (LRT) statistic that is distributed as a χ^2 random variable with $df = 2$. A significant LRT statistic is taken as evidence against the null hypothesis that the studied indicator is measurement invariant. Equivalently, a Wald test statistic can be used. A Wald test is asymptotically equivalent to the LRT (Buse, 1982) but advantageously only requires estimating the unconstrained model, not any constrained models.

Multiple single-group modeling approaches, including RFA and MNLFA, have been proposed for the purpose of assessing measurement invariance. These approaches share the same general form (Equation 4.3), but differ in the way the background variable V is modeled and $\boldsymbol{b}$ and $\boldsymbol{c}$ are estimated. We will discuss the RFA and MNLFA approaches in the following paragraphs. First, we will describe RFA followed by a description of MNLFA, because an RFA model can be seen as a restrictive MNLFA model.

4.2.1 Restricted Factor Analysis

In RFA, the variable V—across which measurement invariance is potentially violated—is added to the single-group model as an exogenous variable that covaries with the common factor T. This covariance captures how common-factor means differ across V. MIMIC models are statistically equivalent to RFA models but include a direct effect of V on the

common factor instead of a covariance. This direct effect can readily be interpreted as the difference in common-factor means for each 1-unit increase in V.

Measurement invariance is evaluated in an RFA model by means of direct and interaction effects of the background variable V on the indicators X. In order to assess scalar invariance, the elements in $\boldsymbol{b}$ are modeled as direct effects of V on X. A nonzero effect of V on X implies that the observed measure depends on V even when holding the common factor constant (i.e., the indicator's intercept τ differs with V, controlling for T). In order to assess metric invariance, the elements in $\boldsymbol{c}$ are modeled as interaction effects between T and V (i.e., $T \times V$) on X. A nonzero interaction effect implies that the magnitude of DIF varies with T (i.e., the indicator's factor loading λ differs with V).

Using maximum likelihood to estimate RFA poses a challenge to testing metric invariance because estimating $\boldsymbol{c}$—the $T \times V$ interaction effects on X—would require modeling the product between V (which could be observed or latent) and the latent common factor T. LMS provides an analytical solution to estimate these interaction effects in RFA models (Barendse et al., 2010; Woods & Grimm, 2011), and Kolbe & Jorgensen (2018, 2019) proposed the PI method as a more widely available alternative. Next, we elaborate on both methods to model interactions in RFA models.

Latent Moderated Structural Equations

LMS is a distributional analytic approach for the estimation of latent interaction effects in structural equation models (A. Klein & Moosbrugger, 2000). With LMS the variable V is modeled as a single-indicator latent variable in the RFA model, which allows $\boldsymbol{c}$ to be estimated as latent interaction effects on the indicators X. The latent interaction effects are estimated by means of a finite mixture of multivariate normal distributions, which takes into account the nonnormality induced by multiplying two normally distributed latent factors. Specifically, the distribution of the observed variables X is regarded as finite mixtures of multiple distributions conditional on the latent variables.

Figure 4.1 shows an RFA model amenable to LMS for assessing measurement invariance with respect to variable V. In this example, T is measured by P indicators denoted X, and V is measured by a single indicator Y. In order for the model to be identified, the factor loading and residual variance of Y are commonly fixed at unity and zero, respectively. Instead of modeling the interaction of T with V as a factor with observed indicators, the LMS approach estimates the interaction effect of $T \times V$ directly using mixture distributions (A. Klein & Moosbrugger, 2000). Therefore, the interaction of T with V is represented in Figure 4.1 by the product $T \times V$ in a dotted circle. Note that associations (i.e., covariances) of the product factor with T and V are not explicitly depicted in Figure 4.1 because they are not estimated, but the estimation implicitly allows those associations exist. A nonzero effect of V on X_p, denoted b_p, implies uniform DIF for indicator p, whereas a nonzero effect of $T \times V$ on X_p, denoted c_p, implies nonuniform DIF for indicator p.

The LMS approach is a full information maximum likelihood approach that assumes

Figure 4.1: An RFA model with LMS for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to assess scalar and metric invariance, respectively.

multivariate normality for all exogenous variables (e.g., the common factors and residuals) in the model. But when V is a categorical variable, this normality assumption is clearly violated. Studies showed that LMS provides efficient estimators when the distributional assumptions are met (A. Klein & Moosbrugger, 2000; Dimitruk et al., 2007), but with nonnormal variables inflated Type I error rates were observed when testing for the significance of a latent interaction effect (A. Klein & Moosbrugger, 2000; A. G. Klein & Muthén, 2007). A violation of multivariate normality can, however, be accounted for by using a robust maximum likelihood estimator (L. K. Muthén & Muthén, 2012). Barendse et al. (2012) provided a description and example syntax of how to apply RFA with LMS in M*plus* (L. K. Muthén & Muthén, 2012).

Product Indicators

The PI method by Kenny & Judd (1984) involves the formation of product indicators that serve as indicators of an ad hoc latent interaction factor representing the interaction between two latent variables. There are various ways to compute the product indicators of the latent interaction factor. Most recently, the double-mean-centering strategy was proposed (Lin et al., 2010). With this strategy, product indicators are built by mean-centering the product terms obtained by multiplying the mean-centered indicators of the associated latent variables. Kolbe & Jorgensen (2018) provided an R (R Core Team, 2018) syntax example of RFA with the PI method using the R packages `lavaan` (Rosseel, 2012) and `semTools` (Jorgensen et al., 2019). Note that an advantage of the PI method is that it can be applied using any standard SEM software because it merely requires calculating
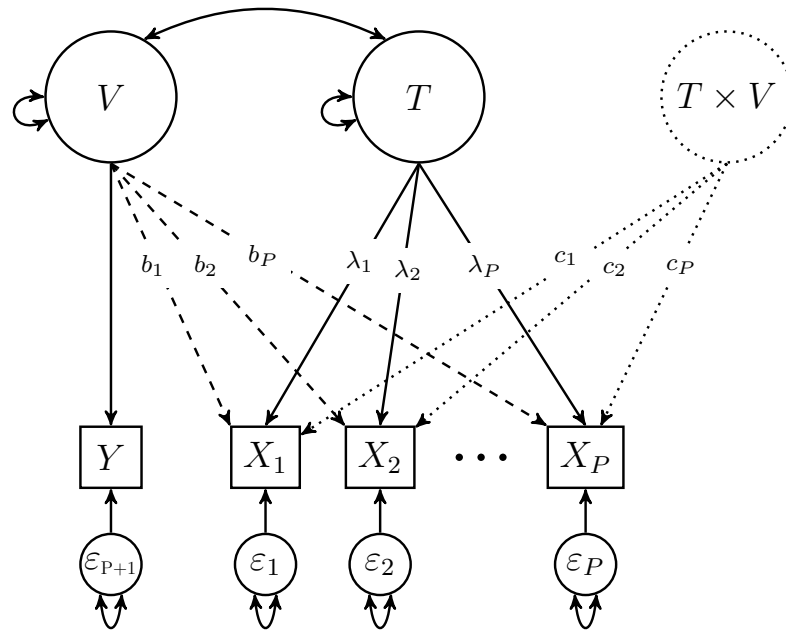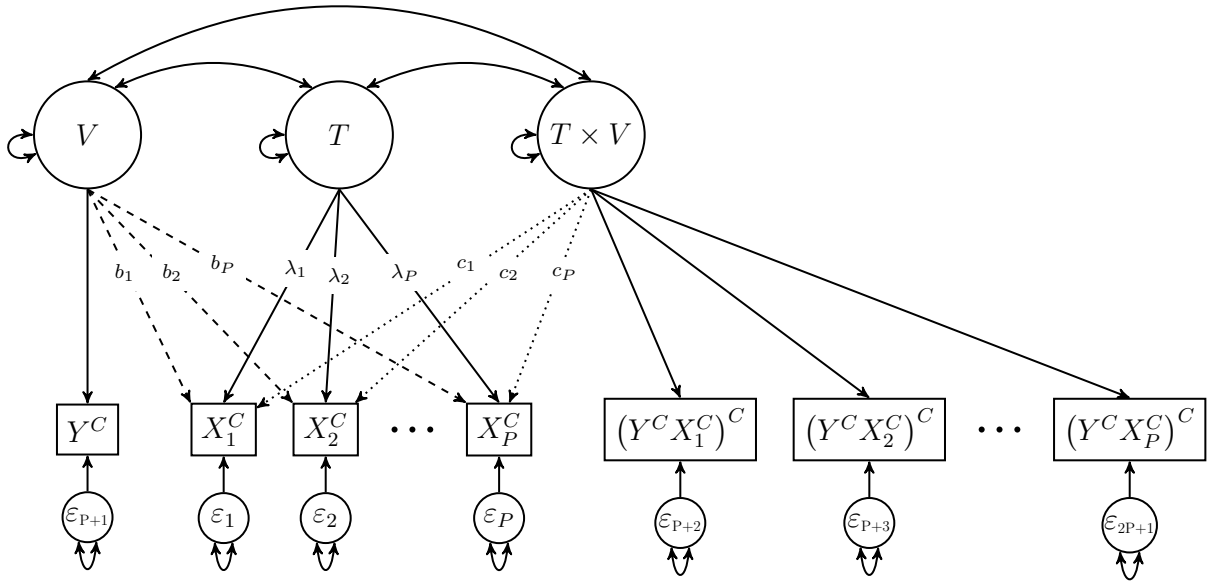
Figure 4.2: An RFA model with PI for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to assess scalar and metric invariance, respectively.

products of indicator scores to be treated as indicators of the latent interaction factor.

Figure 4.2 depicts an RFA model for the assessment of measurement invariance in which the latent interaction factor $T \times V$ is measured by double-mean-centered product indicators. The potential source of a violation V is a latent variable measured by the indicator Y. Similar to LMS, the factor loading and residual variance of Y can be fixed at unity and zero, respectively, in order for the model to be identified. As illustrated in Figure 4.2, the indicators of T and V are mean-centered. The double-mean-centered product indicator of the p-th indicator is denoted $(Y^C \times X_p^C)^C$. Nonzero b and c parameters imply violations of scalar and metric invariance, respectively. Whereas LMS only estimates the covariance between T and V, the PI method additionally allows for the estimation of the covariance between V and $T \times V$ as well as the covariance between T and $T \times V$. The latter covariance will be nonzero only when the common-factor variance differs across levels of V, thus accounting for common-factor heteroskedasticity (see **Appendix I** for details).

The maximum likelihood estimation procedure typically used with the PI method assumes multivariate normality of all indicators in the model (including the product indicators). This assumption is inevitably violated because even products of normal variables are not normally distributed (Jöreskog & Yang-Wallentin, 1996). A robust maximum likelihood estimator can be used to correct for nonnormality (Satorra & Bentler, 2010). Studies have shown that PI methods, including the double-mean-centering strategy, are generally robust against violations of multivariate normality of the product indicators (Marsh et al., 2004; Lin et al., 2010).
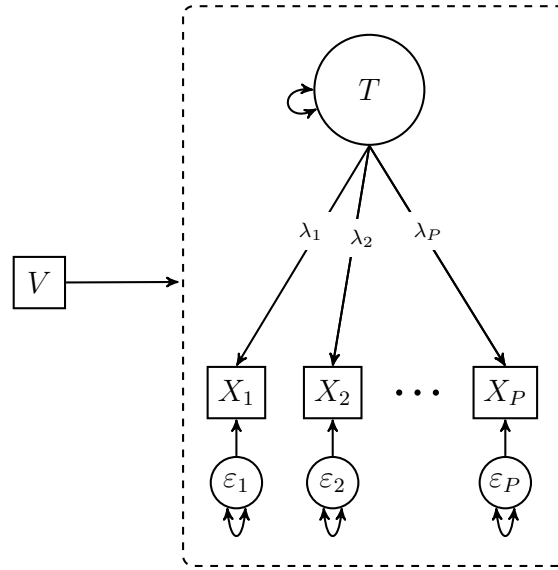
Figure 4.3: An MNLFA model for assessing measurement invariance. The variable V may have an effect on all parameters in the model represented in the dashed border.

4.2.2 Moderated Nonlinear Factor Analysis

The MNLFA approach (Bauer & Hussong, 2009; Bauer, 2017) includes the background variable V in the model only as a moderator variable, whereby parameters can be defined as functions of V. Figure 4.3 illustrates the parameter moderation with the arrow pointing from V to the measurement model for the indicators X. Subject to identification constraints, the variable V may be a predictor of any parameter in the factor analysis model, including the common factor mean and variance, each indicator's intercept and residual variance, and all factor loadings. Thus, no latent interaction is needed.

Measurement invariance can be assessed for each indicator by testing whether V moderates the indicator's intercept τ or factor loading λ. To assess scalar invariance, the vector of intercepts can be written (following from Equation 4.3) as

$$\boldsymbol{\tau}_i = \boldsymbol{\tau}_0 + \boldsymbol{b}v_i, \tag{4.4}$$

where any nonzero element of $\boldsymbol{b}$ indicates a linear change in $\boldsymbol{\tau}$ associated with V (i.e., uniform DIF). Metric invariance can be assessed by expressing factor loadings as

$$\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_0 + \boldsymbol{c}v_i, \tag{4.5}$$

where any nonzero element of $\boldsymbol{c}$ reflects a linear change in $\boldsymbol{\Lambda}$ associated with V (i.e., nonuniform DIF).

In addition to measurement parameters, factor means and variances may also depend on V. For example, the mean of the common factor T can be written as

$$\alpha_i = \alpha_0 + gv_i. \tag{4.6}$$

Here α_0 is the baseline common-factor mean when $v_i = 0$ and g captures the linear effect of V on the common-factor mean. Similarly, the common-factor variance can be expressed as a function of V, but a linear regression model is not suitable for variances because it allows for negative values. Therefore, Bauer & Hussong (2009) proposed to model variances as exponential functions of V. The variance of the common factor T may be written as

$$\psi_i = \psi_0 \exp(hv_i), \tag{4.7}$$

where ψ_0 is the baseline common-factor variance when $v_i = 0$ and h is the effect of V on the common-factor variance. This effect thus captures heteroskedasticity of the common-factor. To model the indicators' residual variances as a function of V, one can adopt the same idea as above, that is,

$$\boldsymbol{\varepsilon}_i = \boldsymbol{\varepsilon}_0 \exp(\boldsymbol{d}v_i), \tag{4.8}$$

where $\boldsymbol{\varepsilon}_0$ is a vector of baseline residual variances and the effects of V on the residual variances are captured by $\boldsymbol{d}$. The baseline coefficients for the common factor α_0 and ψ_0 can be fixed at zero and one, respectively, in order to identify the model in the situation that an anchor indicator's intercept and loading are not constrained to zero and one for identification.

Although MNLFA and RFA differ in the way V is modeled and $\boldsymbol{b}$ and $\boldsymbol{c}$ are estimated, they share the same general model for the indicators X (Equation (4.3)). The MNLFA model is equivalent to the RFA model when only the factor means, indicators' intercepts, and factor loadings are linearly moderated by V. However, the advantage of MNLFA over RFA is that it also allows the common-factor variance and the indicators' residual variances to vary as a function of V. The MNLFA method can thus be conceptualized as an extended RFA model in which variances need not be assumed equal across different levels of V (Bauer, 2017), making it potentially as unrestrictive as multigroup CFA when V is a grouping variable, yet more so because V can also be continuous. Bauer (2017) provided SAS and M*plus* (L. K. Muthén & Muthén, 2012) syntax examples of MNLFA in their supplementary materials. For more details about MNLFA and its precursors, see Neale (1998); Neale et al. (2006); Mehta & Neale (2005); Molenaar, Dolan, Wicherts, & van der Maas (2010); or Purcell (2002).

4.3   Method

We conducted a Monte Carlo simulation study to evaluate the robustness of RFA/LMS, RFA/PI, and MNLFA against violations of the homoskedasticity assumption in the case of categorical and continuous V. The outcomes of interest were Type I error rates and power, which we evaluated for each method under multiple conditions that differed with respect to five design factors:

1. Type of noninvariance: scalar or metric.

2. Total sample size: $N = 100, 200, 500,$ or 1000.

3. Type of V: categorical or continuous.

4. Magnitude and direction of common-factor heteroskedasticity.

5. Magnitude and direction of residual heteroskedasticity.

The levels of the first design factor varied within replications, by assigning different indicators to have different types of noninvariance. We did not vary the magnitude of noninvariance as a design factor because the focus of the current study was not on the impacts of violations of measurement invariance, but on the impacts of different sources of heteroskedasticity on (a) the power to detect violations of measurement invariance and (b) the Type I error rates when indicators have truly invariant measurement parameters. The remaining four design factors were between-replications factors that were fully crossed. For each of these conditions, 1000 replications were generated. The relatively small group sample sizes ($\frac{N}{2}$) were investigated because in such conditions single-group models such as RFA models would be preferred over MGCFA (Oort, 1998), as would be preferable when V is continuous (regardless of sample size).

4.3.1 Data Generation

Data were simulated under different sample sizes using the following data-generating model

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda} t_i + \boldsymbol{b} v_i + \boldsymbol{c} t_i v_i + \boldsymbol{\varepsilon}_i \tag{4.9}$$

where $\boldsymbol{x}_i$ is a vector of 10 continuous indicator scores, t_i is the common-factor score, v_i is the score on the background variable, and $\boldsymbol{\varepsilon}_i$ is a vector of 10 residual scores of subject i. Moreover, the vector $\boldsymbol{\tau}$ includes 10 intercepts set at 0 for all indicators, $\boldsymbol{\Lambda}$ includes 10 common factor loadings set at 0.8 for all indicators, and $\boldsymbol{b}$ and $\boldsymbol{c}$ are vectors of regression coefficients fixed at 0 for all indicators that did not violate measurement invariance.

How we violated invariance and homoskedasticity assumptions in our population model depended on whether V was continuous or categorical. For violations of both common-factor and residual homoskedasticity, we strove to vary the variances such that they ranged from approximately half to double the variance across the range of V, whether that range was across two categories or across two or three standard deviations above and below the mean of V.

Continuous Background Variable

In conditions where the background variable V is a continuous variable, scores on the background variable were drawn from a standard normal distribution $v_i \sim \mathcal{N}(0, 1)$. The common-factor scores t_i were drawn from a normal distribution with a mean equal to v_i and a variance of either 1, $\exp(-0.25 v_i)$, or $\exp(0.25 v_i)$. Hence, there were three levels

of common-factor heteroskedasticity: $h = -0.25$, $h = 0$ (i.e., homoskedasticity), and $h = 0.25$. Figure 4.4 shows the common-factor variances as a function of V for different levels of h. In the two heteroskedastic conditions, the population common-factor variances ranged from 0.61 to 1.65 for $-2 \leq V \leq 2$ and from 0.47 to 2.12 for $-3 \leq V \leq 3$.
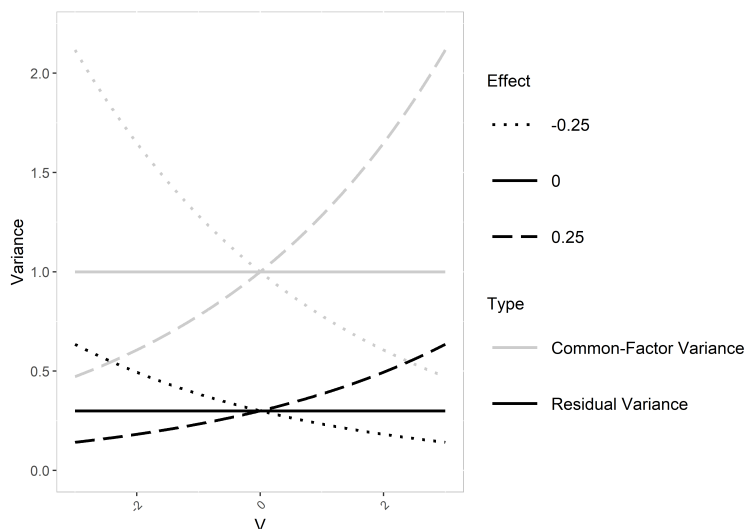
Figure 4.4: The common-factor and residual variances as a function of continuous V.

Residual scores of each indicator were drawn from a normal distribution $\varepsilon_i \sim \mathcal{N}(0, 0.3)$ in conditions with residual homoskedasticity. In order to test the effect of residual heteroskedasticity with respect to a continuous V on the power and Type I error rates, the residuals of one measurement-invariant indicator (Indicator 1) and two indicators that violated measurement invariance (Indicator 2 with uniform DIF and Indicator 4 with nonuniform DIF) were drawn from a normal distribution with a mean of 0 and variance of either 0.3 (in the homoskedastic conditions), $0.3\exp(-0.25v_i)$, or $0.3\exp(0.25v_i)$. This resulted in three levels of residual heteroskedasticity: $d = -0.25$, $d = 0$ (i.e., homoskedasticity), and $d = 0.25$. Figure 4.4 shows the residual variances as a function of V for different levels of d. The residual variances in the population ranged from 0.18 to 0.49 for $-2 \leq V \leq 2$ and from 0.14 to 0.64 for $-3 \leq V \leq 3$ in the two conditions with residual heteroskedasticity.

Uniform DIF was introduced by setting $b = 0.25$ for the second and third indicators, and nonuniform DIF was introduced by setting $c = 0.1$ for the fourth and fifth indicators. These magnitudes reflect small effects of V and $T \times V$ on these indicators (Cohen, 1988). A table with the population parameter values for each indicator is available in **Appendix II**.

Categorical Background Variable

In conditions where the background variable V is a categorical variable, we generated a dummy code that represent group membership. In specific, we chose $v_i = 0$ for the reference group and $v_i = 1$ for the focal group (for more than two groups, multiple

dummy codes would be necessary). The common-factor scores t_i were drawn from a normal distribution with a mean of 0 for the reference group and a mean of -0.5 for the focal group, representing a moderate difference between groups (Kolbe & Jorgensen, 2019). The population common-factor variance in the reference group was equal to 1, whereas the population common-factor variance of the focal group was equal to 0.5, 1, 1.5, or 2. Hence, in total there were four levels of common-factor heteroskedasticity: $h = \ln(0.5)$, $h = 0$ (i.e., homoskedasticity), $h = \ln(1.5)$, and $h = \ln(2)$.

Residual scores of each indicator for the reference group—and all but three indicators in the focal group—were drawn from a normal distribution $\varepsilon_i \sim \mathcal{N}(0, 0.3)$. The residual variances of one measurement-invariant indicator (Indicator 1) and two indicators that violated measurement invariance (Indicator 2 with uniform DIF and Indicator 4 with nonuniform DIF) were 0.15, 0.3, or 0.6 for the focal group, representing three levels of residual heteroskedasticity: $d = \ln(0.15/0.3)$, $d = 0$ (i.e., homoskedasticity), and $d = \ln(0.6/0.3)$.

A violation of scalar invariance of the second and third indicator was introduced by fixing b at 0.5, and a violation of metric invariance of the fourth and fifth indicator was introduced by fixing c at 0.25. These effect sizes reflect small violations of scalar and metric invariance with respect to a categorical V (Barendse et al., 2010).

4.3.2 Analysis

When measurement invariance was examined with RFA, an unconstrained model was fitted in which all elements in $\boldsymbol{b}$ and $\boldsymbol{c}$ were freely estimated, except for the elements corresponding to the ninth and tenth indicator. These indicators were used as anchor indicators to set the scale of the common factor T and were not assessed for measurement invariance[2]. Violations of scalar and metric invariance were examined simultaneously for each of the nonanchor indicators by testing the null hypothesis that the studied indicator p's $b_p = 0$ and $c_p = 0$ using a 2-df Wald test with $\alpha = .05$ level of significance. In order to enable the estimation of the $\boldsymbol{b}$ and $\boldsymbol{c}$ parameters, V was modeled as a single-indicator factor whose factor loading was fixed at unity and residual variance fixed to zero in the RFA models with PI, whereas this residual variance was fixed at a near-zero value of 0.001 in the RFA models with LMS to prevent estimation problems. A robust maximum likelihood estimator was used to account for violations of the normality assumption.

When indicators were assessed for measurement invariance with MNLFA, a measurement model for the common factor T with indicators X was estimated where the common-

---

[2]In the present study, we focus on the inflation of Type I error rates due solely to unmodeled heteroskedasticity, but see Kolbe & Jorgensen (2019) for guidance on empirically selecting anchor indicators and for the impact of contaminated anchor sets on Type I error rates.

factor mean and variance, the residual variances[3], and nonanchor indicators' intercepts and factor loadings are a function of V. Similar to RFA, the ninth and tenth indicator were used as anchor indicators and were not tested for measurement invariance. The common-factor mean and variance for the reference group ($V = 0$) were fixed at zero and one, respectively, for identification. Violations of scalar and metric invariance were examined simultaneously for each indicator by testing the null hypothesis that the effect of V on the indicator's intercept and factor loading is equal to zero, again tested using a 2-df Wald test with $\alpha = .05$ level of significance. A robust maximum likelihood estimator was used with MNLFA to account for nonnormality.

Power and Type I error rates were calculated across all conditions. Power was estimated as the proportion of replications in which Indicator 2 and Indicator 4 (i.e., indicators with uniform and nonuniform DIF, respectively) were correctly flagged as violating measurement invariance. The Type I error rate was estimated as the proportion of replications in which Indicator 1 (i.e., a measurement-invariant indicator) was incorrectly flagged as violating measurement invariance. A 95% Agresti–Coull confidence interval (CI; Agresti & Coull, 1998) around the expected Type I error rate of $\alpha = .05$ was calculated to evaluate whether observed error rates were statistically significantly different from the nominal value (i.e., by checking whether the observed value was in the 95% CI). We considered values inflated > 0.1 as being substantially important (i.e., practical significance).

In addition to the power and Type I error rates, the accuracy and efficiency of the parameter estimates in $\boldsymbol{b}$ and $\boldsymbol{c}$ of the indicators with DIF were evaluated for each method by calculating the relative bias, root mean squared error (RMSE), and coverage rates. The relative bias of the parameter estimate b of Indicator 2 was defined as a percentage using $((\bar{b} - b)/b) \times 100\%$, where $\bar{b}$ is the average parameter estimate across replications and b is the true parameter value. We considered relative bias larger than 5% as substantial bias. Moreover, the RMSE of the parameter estimate b of Indicator 2 was defined as $\sqrt{(\bar{b} - b)/b}$. The coverage rate of the parameter estimate b of Indicator 2 was defined as the proportion of replications in which the 95% confidence interval around the parameter estimate contained the population value b. The relative bias, RMSE, and convergence rates of the parameter estimate c of Indicator 4 were defined similarly.

The power, Type I error rates, relative bias, RMSE, and coverage rates are presented in figures, but tables of these outcome variables are available in **Appendix II**. The RFA/LMS and MNLFA models were fit in M*plus* (version 7; L. K. Muthén & Muthén, 2012) via the `MplusAutomation` package (version 0.7-2; Hallquist & Wiley, 2018), and the RFA/PI models were fit with the R (version 3.4.3; R Core Team, 2018) package `lavaan` (version 0.5-23; Rosseel, 2012), relying on the `semTools` function `indProd()` to calculate

---

[3]This MNLFA specification allows for both types of heteroskedasticity, so it is therefore less restrictive than RFA. When MNLFA does not include effects of V on variances, it would be as restrictive as RFA/LMS. Because the estimation method is so computationally intensive, we did not include such a "homoskedastic MNLFA" in our simulation. We did conduct example analyses applied to real data, available on our Open Science Framework project `https://osf.io/vsp4f/`, which showed that a homoskedastic MNLFA and RFA/LMS yielded very similar results.

double-mean-centered product indicators. All data generation and analysis of results were conducted in R. See our Open Science Framework project `https://osf.io/vsp4f/` for example scripts.

4.4 Results

Before we present the power and Type I error rates, we first elaborate on the convergence rates of the different methods. Detailed convergence rates across conditions are available in **Appenidx II**. Across all methods, we encountered the largest nonconvergence rates for RFA/LMS. The nonconvergence rates when V was continuous decreased with sample size. In the smallest sample-size conditions the percentages of nonconvergence ranged from 0.10 to 4.80, whereas in the largest sample-size condition the RFA/LMS model always converged.

The nonconvergence rates were substantially larger for RFA/LMS when V was a categorical variable. On average across all conditions with a categorical V, the RFA/LMS model did not converge in 16.64% of all replications. The largest nonconvergence rates were observed in conditions in which the common-factor variance of the focal group was larger than the common-factor variance of the reference group. All replications with nonconvergence were excluded from the analysis for RFA/LMS, because in such replications, measurement invariance could not be assessed with this method.

The MNLFA method only once produced convergence problems. Similar to RFA/LMS, this replication could not be included in the analysis for MNLFA. The RFA/PI models converged for every replication in each condition. Because in some conditions the results for RFA/LMS were based on a notably smaller number of replications compared to RFA/PI and MNLFA, the validity of a comparison between the methods could be questioned. In a comparable study, Kolbe & Jorgensen (2019) showed that using a smaller subset of replications for RFA/LMS does not affect the pattern of the results. Hence, below we present the results based on all available converged replications in each condition.

4.4.1 Continuous Background Variable

Power and Type I Error Rates

The power to detect violations of metric invariance using each method across conditions with a continuous variable V is presented in Figure 4.5. Because for scalar invariance the differences across the methods were quite negligible, we only include a figure for the power to detect scalar invariance in **Appendix III**. For each of the methods, power to detect violations of both scalar and metric invariance increased with sample size and was effectively 1.00 in all conditions with a sample size of $N \geq 500$. More apparent differences in power were observed when $N = 100$ or $N = 200$. The RFA/LMS method generally obtained higher power to detect violations of metric invariance than RFA/PI and MNLFA in conditions with a positive effect of V on the common-factor variance (but at the expense

Figure 4.5: The power to detect a violation of metric invariance of Indicator 4 (i.e., $c_4 \neq 0$) of each method across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

of inflated Type I error rates), and lower power to detect violations of metric invariance than RFA/PI and MNLFA when this effect was negative. Residual heteroskedasticity did not seem to substantially affect the power of the methods.

Figure 4.6 illustrates the Type I error rates of each method in conditions with a continuous variable V. The light gray region from .01 to .10 represents a region of practical equivalence (ROPE), outside of which are substantially inflated error rates. The darker gray region is the Agresti–Coull 95% CI around $\alpha = .05$, values inside of which are not statistically significantly different from the nominal level. When $h = 0$ (common-factor homoskedasticity), Type I error rates were comparable across the three methods and decreased with sample size. In general, Type I error rates in these conditions were only substantially inflated when $N = 100$. Residual heteroskedasticity hardly affected the Type I error rates of any of the methods in conditions where $h = 0$.

In conditions with common-factor heteroskedasticity (i.e., $h = -0.25$ or 0.25), the Type I error rates were substantially different across the methods. In almost all conditions, the RFA/LMS method obtained the most inflated Type I error rates compared to the other methods. Especially when the effects of V on the common-factor variance and residual variances were in similar directions (e.g., $h = 0.25$ and $d = 0.25$), large inflation of the error rates of RFA/LMS was observed, and the inflation was exacerbated in larger
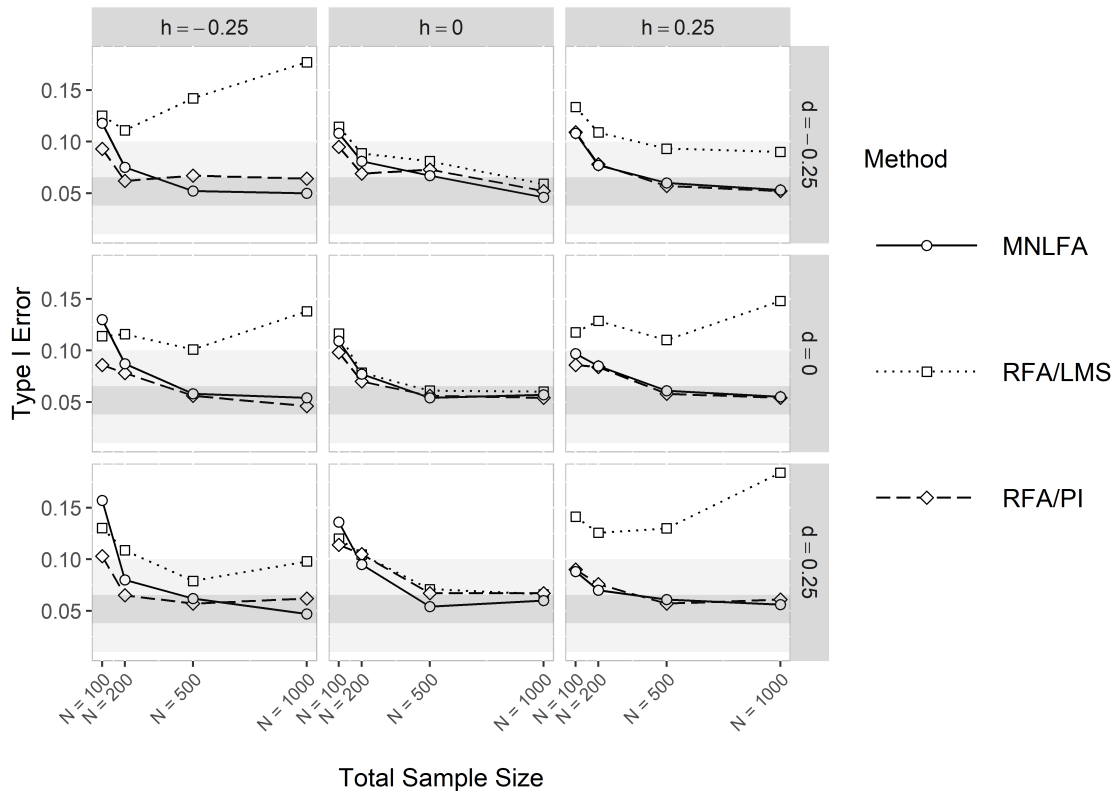
Figure 4.6: The Type I error rates for Indicator 1 of each method across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

samples. In contrast, when the effects on the variances were in opposite directions (e.g., $h = -0.25$ and $d = 0.25$), the Type I error rates of RFA/LMS were less inflated, but almost always remained higher than for other methods. The RFA/PI and MNLFA Type I error rates were not substantially affected by combined common-factor and residual heteroskedasticity. Overall, MNLFA obtained error rates closer to .05 than other methods.

Relative Bias of DIF Estimates

Figures of the relative bias of the b and c parameter estimates across all conditions with a continuous V can be found in **Appendix III**. The relative bias of the parameter estimate b was negligible for RFA/PI and generally acceptable (i.e., smaller than 5%) for RFA/LMS and MNLFA. Larger differences between the methods were observed for the relative bias of the c parameter estimates. Overall, MNLFA obtained the least biased parameter estimates c. The relative bias of this method was always below 5%, except in some conditions where $N = 100$. The RFA/PI and RFA/LMS methods substantially overestimated c in all conditions. The relative bias in c produced by RFA/PI ranged from 23.66% to 26.67% and seemed unaffected by sample size and common-factor and residual heteroskedasticity. The RFA/LMS method obtained the most biased parameter estimates c, with relative bias ranging from 21.93% to 59.56%. The relative bias of this method was largest when

h was positive.

RMSE of DIF Estimates

Figures of the RMSE of the b and c parameter estimates across all conditions with a continuous V can be found in **Appendix III**. The differences between the methods with respect to RMSE of the parameter estimate b were relatively small in all conditions with a continuous V. Overall, the RMSE of the parameter estimate b decreased with sample size but seemed unaffected by common-factor and residual heteroskedasticity. The only conditions in which MNLFA produced a substantially higher RMSE than the other methods were conditions in which $h = -0.25$ and $N = 100$. With respect to parameter estimate c, differences in the RMSE across the methods were observed more frequently. In general, MNLFA obtained the lowest RMSE of the parameter estimate c, followed by RFA/PI. In almost all conditions, RFA/LMS obtained the highest RMSE for the parameter estimate c.

Coverage Rates of DIF Estimates

Figures of the coverage rates of the b and c parameter estimates across all conditions with a continuous V can be found in **Appendix III**. Overall, all methods showed acceptable coverage rates (always $> .90$) for the parameter estimate b. The RFA/PI method obtained coverage rates closest to .95 for b, followed by MNLFA. The coverage rates of RFA/LMS for b were slightly smaller compared to other methods. Different patterns were observed for the coverage rates of the parameter estimate c. Whereas MNLFA frequently obtained coverage rates above .90 for c, RFA/PI and RFA/LMS frequently obtained unacceptable coverage rates. For both methods, the coverage rates for c decreased with N and h. The RFA/LMS method obtained the lowest coverage rates for the parameter estimate c. The lowest coverage rate of .07 was obtained when $h = 0.25$, $d = -0.25$, and $N = 1000$.

4.4.2 Categorical Background Variable

Power and Type I Error Rates

Again, power showed nearly no difference between methods for detecting violations of scalar invariance with respect to a categorical V, so a figure is included only in **Appendix III**. For each of the methods, the power to detect violations of measurement invariance increased as a function of sample size. The power to detect violations of scalar invariance when $N = 100$ ranged from .83 to .98, where a negative effect on the residual variance led to higher power and a positive effect on the residual variance led to lower power for each of the methods. In the other sample-size conditions, the power to detect violations of scalar invariance was generally 1.00. Hence, the methods performed similarly well with respect to detecting violations of scalar invariance.
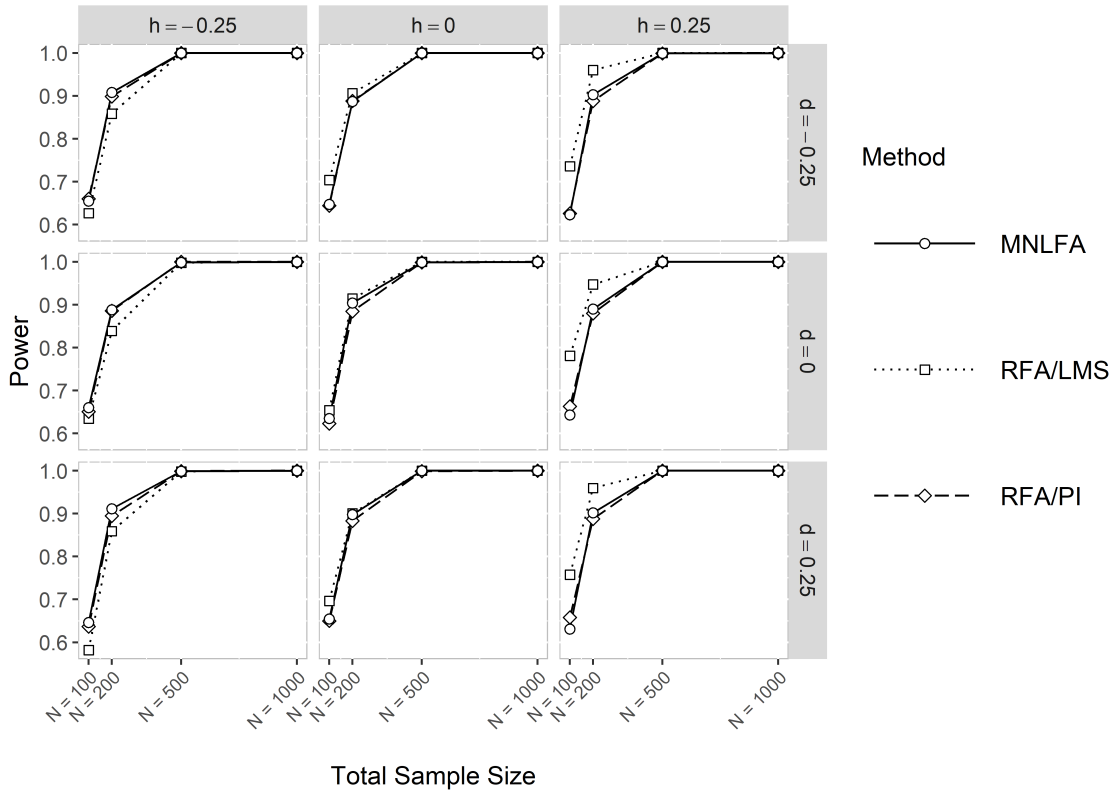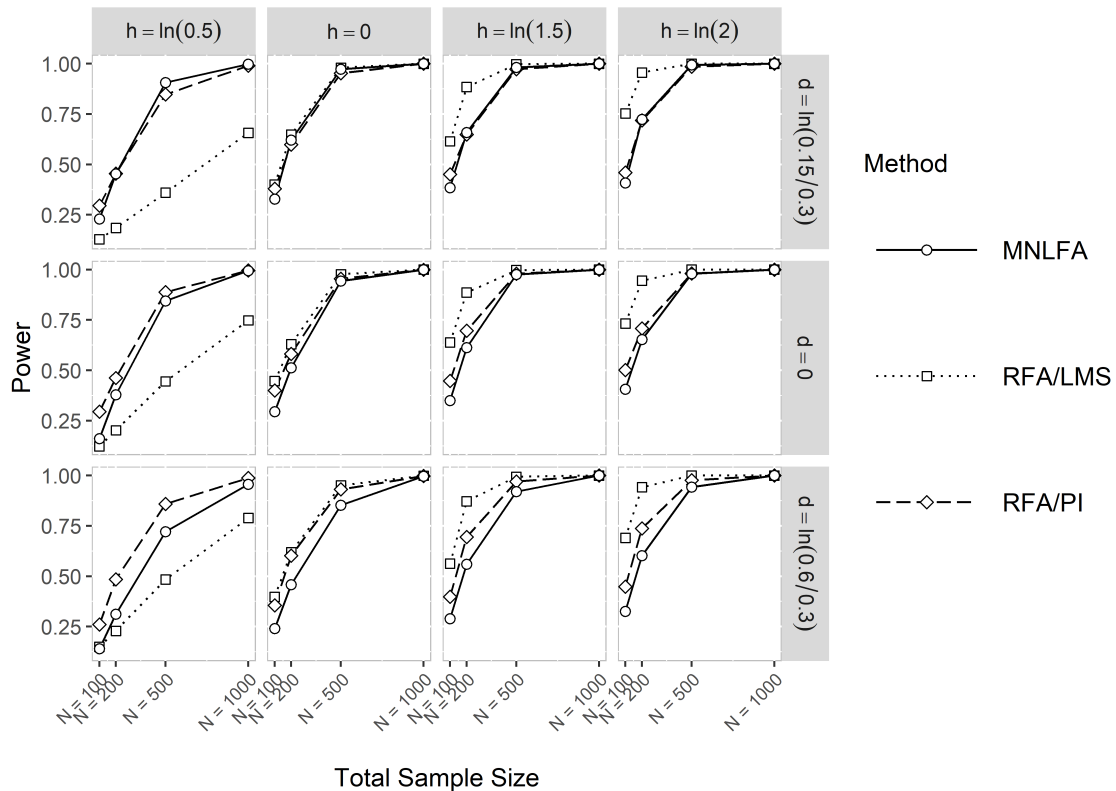
Figure 4.7: The power to detect a violation of metric invariance of Indicator 4 (i.e., $c_4 \neq 0$) of each method across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure 4.7 shows the power of methods to detect violations of metric invariance. In conditions with equal common-factor variances across groups (i.e., $h = 0$), RFA/LMS and RFA/PI obtained slightly higher power than MNLFA. Moreover, RFA/LMS outperformed RFA/PI and MNLFA when the focal group had a larger common-factor variance than the reference group (i.e., $h = \ln(1.5)$ or $h = \ln(2)$), but performed substantially worse when the focal group's common-factor variance was smaller (i.e., $h = \ln(0.5)$).

The Type I error rates across all conditions with a categorical V are illustrated in Figure 4.8. Note that we specified y-axis limits of 0 and .15 in order to make details more visible, at the expense of plotting a few extremely inflated values for RFA/LMS outside the plot range. Type I error rates of all methods under common-factor homoskedasticity were close to the nominal .05, within the ROPE [.01–.10]. The majority of MNLFA's Type I error rates were not significantly inflated, whereas RFA/LMS and RFA/PI had statistically significant error, particularly under residual heteroskedasticity. However, RFA/PI's error rates were not substantially inflated under any conditions (i.e., the Type I error rates were almost always $< .10$).

In contrast, the RFA/LMS method obtained severely inflated Type I error rates under common-factor heteroskedasticity, so severe that many conditions have error rates beyond the y-axis limits (see **Appendix II** for exact error rates). This inflation was smallest
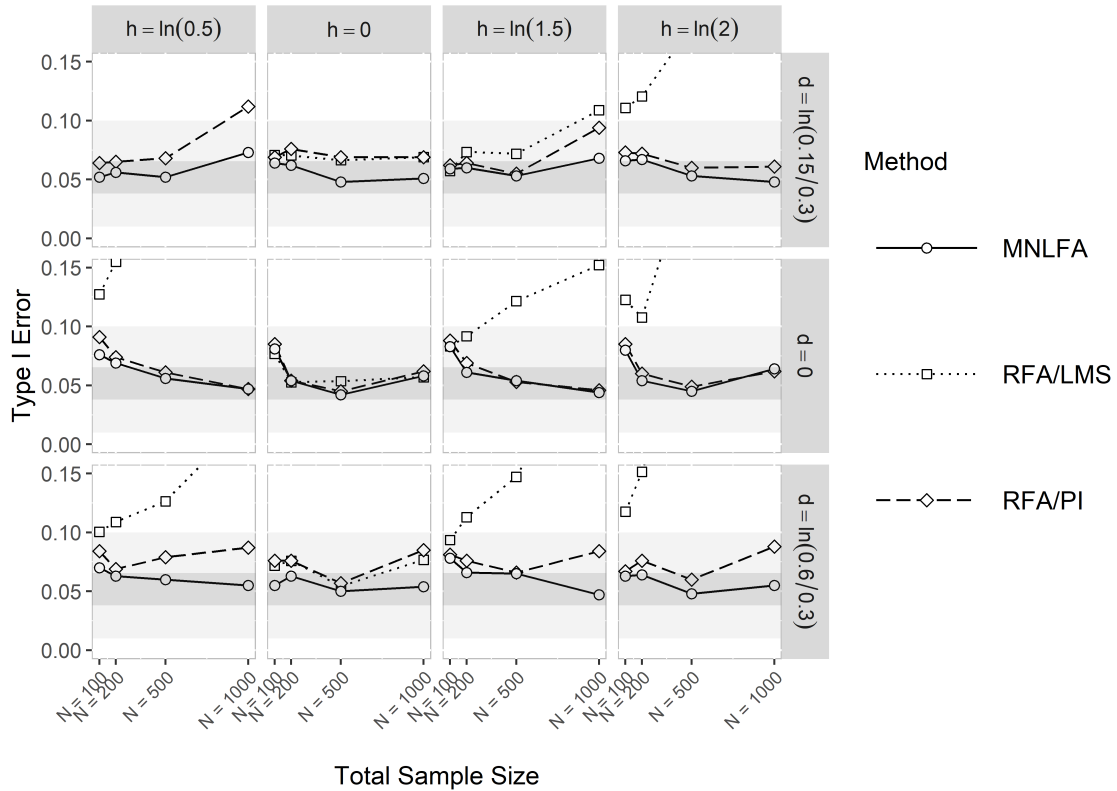
Figure 4.8: The Type I error rates for Indicator 1 of each method across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance. The y axis stops at .15 in order to allow for a detailed comparison of methods with (nearly) nominal error rates, but note that it prevents plotting some extremely inflated error rates in certain conditions of RFA/LMS.

when the effects of V on the common-factor and residual variances were in opposite directions and was largest when these effects were in similar directions. For example, when $h = \ln(0.5)$, $d = \ln(0.15/0.3)$, and $N = 1000$, RFA/LMS obtained a Type I error rate of .87. Though not practically significant, inflation of the Type I error rates of RFA/PI was observed mainly when h and d were both nonzero. The Type I error rates of MNLFA were not substantially affected by common-factor or residual heteroskedasticity.

Relative Bias of DIF Estimates

Figures of the relative bias of the b and c parameter estimates across all conditions with a categorical V can be found in **Appendix III**. The observed patterns were similar to those in conditions with a continuous V. Each method obtained negligible relative bias (i.e., smaller than 5%) of the parameter estimate b, whereas only MNLFA obtained negligible relative bias of the parameter estimate c. Again, the parameter estimates c obtained by RFA/PI and RFA/LMS were substantially biased. The RFA/PI method consistently overestimated c, while RFA/LMS underestimated c in conditions with a negative effect

on the common-factor variance and overestimated c in conditions with a positive effect on the common-factor variance. The relative bias obtained by RFA/LMS was largest when h and d were in similar directions. In contrast, this method generally obtained acceptable relative bias (i.e., smaller than 5%) in homoskedastic conditions.

RMSE of DIF Estimates

Figures of the RMSE of the b and c parameter estimates across all conditions with a categorical V can be found in **Appendix III**. The differences across the methods with respect to RMSE of the parameter estimate b were negligible in all conditions with a categorical V. The differences in RMSE were more apparent for the parameter estimate c. MNLFA obtained the lowest RMSE of the parameter estimate c in almost all heteroskedastic conditions (i.e., $h \neq 0$ or $d \neq 0$). The RFA/PI method generally obtained the second-lowest RMSE of the parameter estimate c when h was positive, but obtained the highest RMSE when h was negative and d was positive. In the conditions with common-factor homoskedasticity (i.e., $h = 0$), MNLFA and RFA/LMS obtained slightly lower RMSE for the parameter estimate c than RFA/PI.

Coverage Rates of DIF Estimates

Figures of the coverage rates of the b and c parameter estimates across all conditions with a categorical V can be found in **Appendix III**. Similar to conditions with a continuous V, all methods obtained acceptable coverage rates for the parameter estimate b (always $> .90$), and only MNLFA obtained acceptable coverage rates for the parameter estimate c in all conditions. The RFA/PI and RFA/LMS methods performed substantially worse than MNLFA with respect to the coverage rates of parameter estimate c. For RFA/PI, the coverage rates for c where acceptable (i.e., larger than .80) in smaller sample-size conditions, but were frequently unacceptable (i.e., smaller than .80) when $N = 500$ or 1000. The RFA/LMS performed worst with respect to the coverage rates for the parameter estimate c, especially in conditions with common-factor heteroskedasticity. In these conditions, the coverage rate of RFA/LMS was .01 at its lowest.

4.4.3   Supplemental Simulation Study

To further investigate the relative robustness of MNLFA and RFA/PI to heteroskedasticity across a wider array of conditions, we conducted additional simulations within the following condition from the original simulation study: a grouping variable V, a total sample size of 200, a factor variance in the focal group of 1.5 (representing common-factor heteroskedasticity), and residual variances of the indicators with unequal residual variances in the focal group of 0.15 (representing residual heteroskedasticity). Within this condition, we fully crossed three additional factors: the total number of indicators (10 or 20), the percentage of indicators violating measurement invariance (40% or 80%), and the percentage of indicators violating residual homoskedasticity (30% or 90%). We generated

data using the same procedure as in the first simulation, and we recorded the effect of these new factors on power and Type I error rates.

The results of the supplemental simulations can be found in **Appendix II**. With all other conditions of the simulation study being equal, the results are comparable to the results of the original simulation study: (a) RFA/LMS is not robust against violations of common-factor and residual homoskedasticity, (b) MNLFA maintains Type I error rates quite well across all conditions, and (c) so does RFA/PI, although not quite so well as MNLFA. Moreover, none of the additional manipulated factors substantially affected the power or Type I error rates. The total number of indicators and the percentage of indicators that violate residual homoskedasticity only led to a minor difference in power and Type I error. Moreover, the percentage of indicators that violate measurement invariance did not seem to affect the power and Type I error at all.

4.5 Discussion

This study addressed the impact of heteroskedasticity on assessing measurement invariance with respect to categorical and continuous observed background variables in single-group models. A common single-group method to assess measurement invariance is RFA (or MIMIC). Previous studies showed that RFA has high power to detect violations of measurement invariance, but severely inflated Type I error rates have also been observed (Woods & Grimm, 2011; Barendse et al., 2010, 2012; Kolbe & Jorgensen, 2019). Most recently, MNLFA was introduced as a single-group method to assess measurement invariance (Bauer, 2017). MNLFA is more flexible than RFA because the former can allow common-factor and residual variances to differ across V. In this study, we examined how the power and Type I error rates of RFA and MNLFA varied as a function of differences in common-factor variances and residual variances with respect to V. Specifically, we compared the performance of RFA/LMS, RFA/PI, and MNLFA under conditions of common-factor and residual homoskedasticity and heteroskedasticity, providing the first empirical evaluation of MNLFA since it was proposed for testing measurement invariance (Bauer, 2017).

In accordance with previous research (Chun et al., 2016; Harpole, 2015), we found that the Type I error rates obtained by RFA/LMS substantially increased as a function of common-factor heteroskedasticity with respect to a categorical V. Whereas in conditions with equal common-factor variances the Type I error rates were only occasionally and slightly inflated, the error rates were severely inflated when common-factor variances differed across groups. The inflation of the Type I error rates obtained by RFA/LMS was largest when the effect of the categorical V on the common-factor variance and residual variances was in similar directions. We observed comparable patterns but less severely inflated Type I error rates of RFA/LMS in conditions with a continuous V. Although the range of differences in variances was comparable between categorical- and continuous-V conditions, differences can be considered more severe in the categorical conditions because

all cases are drawn from distributions with variances at one extreme or another, rather than variances along a continuum between those extremes.

Overall, the results of the present study suggest that RFA/LMS is not robust to common-factor or residual heteroskedasticity. As in previous research (Kolbe & Jorgensen, 2019), we observed a large percentage of nonconvergence for RFA/LMS, especially when V is a categorical variable. This is an important practical limitation of LMS because it may prevent researchers from being able to infer whether indicators are measurement invariant with respect to V.

Following previous research findings (Kolbe & Jorgensen, 2019), we expected no impact of common-factor heteroskedasticity for RFA/PI. The results of this study indeed suggest that RFA/PI is robust against violations of the common-factor homoskedasticity assumption. This observation coincides with the mathematical proof in **Appendix I**, showing that the covariance between the common factor T and the interaction factor $T \times V$—which is estimable with RFA/PI but not with RFA/LMS—indirectly captures information about the difference in common-factor variances across different levels of V. Similar to the RFA/LMS model, the RFA/PI model does assume residual homoskedasticity. The Type I error rates of RFA/PI were slightly inflated by residual heteroskedasticity across a categorical V. When V was a continuous variable, similar patterns were observed but the Type I error rates were less severely inflated.

In contrast to RFA, the MNLFA method does not need to assume homoskedastic common factors or residuals across V. This is because in MNLFA models each parameter including common-factor variances and residual variances of the indicators may be moderated by V. We therefore expected that the Type I error rates were unaffected by heteroskedasticity. In accordance with our expectations, the magnitude of the difference in common-factor and residual variances did not seem to have any impact on the Type I error rates of MNLFA. Both in conditions with a categorical and continuous V, the Type I error rates of this method were rarely inflated. Hence, the results of this study suggest that MNLFA can better minimize Type I error rates than RFA when residual variances differ with respect to V. The present study only investigated a limited number of conditions that varied with the magnitude of heteroskedasticity and sample size. It would be valuable to further investigate the performance of MNLFA as a tool for measurement invariance assessment under other conditions, such as different numbers of indicators, multiple variables V (including multiple dummy codes for a single categorical variable), unbalanced samples, or nonlinear moderating effects.

It is worth noting that despite the advantages of MNLFA, it is only implemented in M*plus* (L. K. Muthén & Muthén, 2012) and OpenMx (Boker et al., 2011); although it could easily be implemented in general Bayesian software, it is not yet available in other dedicated SEM software packages. Of the methods considered in this study, only RFA/PI can be implemented in any SEM program. Because we have shown RFA/PI to be practically robust to heteroskedasticity (i.e., minimally inflated error rates), we can recommend its use to researchers without access to M*plus* or when MGCFA is underpowered (due to

small N) or inappropriate (continuous V).

In addition to the Type I error rates, we examined the power of each method to detect violations of measurement invariance. Because the Type I error rates of RFA/LMS were severely inflated in conditions with heteroskedasticity, we advise against comparing its power to the other methods. However, a valid comparison between MNLFA and RFA/PI can be made. In each of the conditions, the power to detect violations of scalar invariance was generally comparable across these two methods. A larger difference between the methods occurred for the power to detect violations of metric invariance. These differences were most apparent in smaller samples, where RFA/PI was generally more powerful than MNLFA. This method could therefore be preferred over MNLFA in small samples.

An examination of the accuracy and efficiency of DIF parameter estimates revealed large differences between the methods. MNLFA performed substantially better than RFA/PI and RFA/LMS with respect to relative bias, RMSE, and coverage rates of nonuniform DIF estimations (i.e., $\hat{c}$). Both RFA/PI and RFA/LMS yielded biased estimates and low coverage rates for the effects that reflect a violation of metric invariance. The practical impact seems especially problematic for RFA/LMS because of its severely inflated Type I error rates.

In addition to RFA and MNLFA, many other methods for assessing measurement invariance have recently been proposed, including SEM trees (Brandmaier et al., 2013). SEM trees allow for the detection of heterogeneity with respect to continuous or categorical variables by recursively partitioning the data into subsets with significantly different SEM-parameter estimates. Although simulation studies showed that SEM trees are generally able to correctly partition the data into subsets with different parameter estimates (Usami et al., 2017, 2019) and detect uniform DIF in an IRT framework (Tutz & Berger, 2016; Strobl et al., 2015), these methods have only been shown to be effective in large samples, which is a common result for machine-learning algorithms in general. Other methods for the assessment of measurement invariance worth investigating are local SEM (LSEM; Hildebrandt et al., 2016), heteroskedastic latent trait models (Molenaar et al., 2012; Molenaar, 2015; Molenaar et al., 2011; Molenaar, Dolan, & Verhelst, 2010), and stochastic process-based testing (Merkle & Zeileis, 2013; Merkle et al., 2014). An advantage of LSEM and heteroskedastic latent trait models is that these methods can easily be adapted for binary and ordinal indicators; stochastic process-based testing can too, but it is more suitable for ordinal background variables V.

Although indicators in the present study are assumed to be continuous, MNLFA and RFA/LMS can also handle binary and ordinal indicators (see Bauer, 2017; Woods & Grimm, 2011). A generalization of RFA/PI for binary and ordinal indicators is less straightforward. For example, if both the indicators of T and the background variable V are ordinal, the indicators of the latent interaction factor $T \times V$ are products of ordinal indicators. This brings up the question of how products of ordinal indicators can be interpreted (e.g., what is the measurement level of such indicators?). In a recent simulation study, Lodder et al. (2019) evaluated the performance of the PI method in conditions with

ordinal data in a more general context of latent interactions among common factors. The results of their simulation study showed that treating the product indicators as continuous performs at least as well as treating them as ordinal in terms of power, Type I error, and estimation bias. Given that the use of product indicators for the specific purpose of measurement invariance assessment with ordinal data is yet unexplored, much more research is needed to evaluate its performance.

The present study illuminated the impact of unmodeled heteroskedasticity on assessing measurement invariance using single-group models. In the presence of heteroskedastic common factors or residuals, we advise against using the LMS method in RFA models because of severely inflated Type I error rates. RFA/PI and MNLFA are quite robust to heteroskedasticity because these models (at least partially) account for it. Further evaluation of MNLFA for assessing measurement invariance is warranted.

Chapter 5

Assessing Measurement Invariance with Moderated Nonlinear Factor Analysis

Abstract

Assessing measurement invariance is an important step in establishing a meaningful comparison of measurements of a latent construct across individuals or groups. Most recently, moderated nonlinear factor analysis (MNLFA) has been proposed as a method to assess measurement invariance. In MNLFA models, measurement invariance is examined in a single-group confirmatory factor analysis model by means of parameter moderation. The advantages of MNLFA over other methods are that it (1) accommodates the assessment of measurement invariance across multiple continuous and categorical background variables and (2) accounts for heteroskedasticity by allowing the factor and residual variances to differ as a function of the background variables. In this chapter, we aim to make MNLFA more accessible to researchers without access to commercial structural equation modeling software by demonstrating how this method can be applied with the open-source R package `OpenMx`.

5.1 Introduction

The field of psychology is dominated by the use of questionnaires or tests that measure latent constructs like cognition, attitude, and personality. The observed scores derived from these measurement instruments are often used for decisions about, for example, which applicant is best suited for a position, whether a child is suffering from an anxiety disorder, or to what extent an intervention has improved a patient's well-being. Given the importance of these decisions, it is crucial that the construct is measured equivalently across individuals, groups, or over time. This condition is often referred to as measurement invariance (Meredith, 1993). If measurement invariance does not hold, observed differences between individuals or groups may reflect differential measurement and not true differences on the latent construct of interest. Assessing measurement invariance has thus become an important step in psychometric research and applications. Measurement invariance is commonly assessed by fitting a latent variable model to the data obtained from questionnaires or tests with multiple items. A measurement instrument is invariant if for each item the observed score at any given level of the latent construct is not affected by any background variables. For example, the expected observed score of an item on a social anxiety questionnaire should be the same across boys and girls who have the same level of social anxiety. An item that fails to meet this condition of measurement invariance indicates differential item (or indicator) functioning (DIF; Mellenbergh, 1989).

Measurement invariance is often examined with respect to a grouping variable using multiple-group confirmatory factor analysis (MGCFA; Vandenberg & Lance, 2000). In MGCFA, a confirmatory factor analysis (CFA) model is estimated for each group and invariance constraints are imposed on the parameter estimates (i.e., factor loadings, intercepts, and residual variances) in order to assess increasingly restrictive levels of measurement invariance. If an omnibus null hypothesis (H_0) of a specific level of measurement invariance is rejected, follow-up tests can be performed in order to explore which items function differently across the groups. One strength of MGCFA is that all parameters, such as common-factor means and variances, can differ between groups. However, the accompanying limitation is that MGCFA is only designed to assess measurement invariance across a single categorical background variable. In light of this limitation, Bauer (2017) proposed moderated nonlinear factor analysis (MNLFA) as a more flexible alternative to MGCFA. The MNLFA approach examines measurement invariance in a single-group CFA model by means of parameter moderation. In contrast to MGCFA, the MNLFA approach accommodates the assessment of measurement invariance across multiple continuous and categorical background variables simultaneously.

Several empirical validation and simulation studies have examined the performance of MNLFA as a measurement invariance assessment tool (Bauer, 2017; Bauer et al., 2020; Kolbe et al., 2021). These studies showed that the method performs well both in small and large samples and with categorical and continuous data. Specifically, the results of the studies indicated that MNLFA yields unbiased parameter estimates, minimizes Type

I error rates, and has acceptable to high power. This method effectively detects true violations of measurement invariance and avoids detecting negligible violations, particularly when using regularization (Bauer et al., 2020). Given its flexibility and good performance, MNLFA seems to be a promising method for evaluating measurement invariance for a great variety of researchers. But until now, the majority of available guidelines on how to perform this method involve commercial structural equation modeling (SEM) software (i.e., M*plus* and SAS; see Bauer, 2017). There seems to be a lack of documentation on applying this method in open-source SEM software, which is more widely available for the global community of researchers. Performing MNLFA for measurement invariance assessment may therefore not be straightforward for researchers without access to M*plus* or SAS.

This chapter presents a tutorial on assessing measurement invariance through MNLFA with the R (R Core Team, 2021) package `OpenMx` (Boker et al., 2011)[1]. Our aim is to make MNLFA more accessible for any researcher by providing a detailed guideline on performing the method in this open-source SEM software. We will demonstrate MNLFA with a two-factor model and two background variables (categorical and continuous), but it can easily be applied to single-factor models, structural models, or models with fewer or more background variables. In the next section, we introduce the concept of measurement invariance. We then provide a brief explanation of the MNLFA approach, followed by a step-by-step guide for assessing measurement invariance with this method using `OpenMx`. Lastly, we offer concluding remarks regarding the use of MNLFA and address latest developments for measurement invariance assessment with open-source software packages.

5.2 Background

5.2.1 Measurement Invariance

In this section, we give a formal definition of measurement invariance that will be useful for understanding how it can be assessed with the MNLFA approach. Measurement invariance is said to hold if the distribution of observed item responses is not affected by any variables other than the latent constructs of interest (Mellenbergh, 1989). Given that the latent constructs are known, the definition can be mathematically expressed as

$$f_1(X|T,V) = f_2(X|T) \tag{5.1}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ denote the probability distributions of a set of observed variables X (e.g., items) measuring the latent construct T and V is a set of one or more background

---

[1]The open-source R packages `mnlfa` (Robitzsch, 2019), `GPCMlasso` (Schauberger, 2021), and `regDIF` (Belzak, 2021) can be used to estimate MNLFA models as well, but their implementations are currently limited to unidimensional item response theory (IRT) models. Note that MNLFA can also be estimated with open-source Bayesian software like Stan (Stan Development Team, 2021). An R script demonstrating how to estimate MNLFA with Stan is available on our Open Science Framework project https://osf.io/6cyxt/.

variables such as age or gender. This mathematical expression states that measurement invariance holds if the distribution of X depends only on the latent construct T and is invariant with respect to background variable(s) V. Note that the condition of measurement invariance still allows for a relationship between X and V, but it does preclude a direct effect of V on the distribution of X other than through its influence on T.

If measurement invariance does not hold, the observed item responses depend not only on the latent construct but also directly on the background variable(s). In other words, a violation of measurement invariance indicates that the relationship between the observed item responses and the latent construct differs as a function of the background variable(s). An item that violates measurement invariance is said to show DIF. A distinction can be made between full and partial invariance, where full invariance implies that all items of a test or questionnaire are measurement invariant and partial invariance implies that measurement invariance only holds for a subset of items and some items show DIF. Under partial invariance, groups or individuals can still be validly compared on the measurement of the latent construct as long as DIF is correctly detected and modeled.

Different levels of measurement invariance for CFA models have been defined (Meredith, 1993; Steenkamp & Baumgartner, 1998; Horn & McArdle, 1992). Consider a multidimensional factor model in which the item responses X serve as indicators for multiple common factors T. This model can be specified as

$$\boldsymbol{x}_i = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{t}_i + \boldsymbol{\varepsilon}_i. \tag{5.2}$$

Here $\boldsymbol{x}_i$ is a $P \times 1$ vector of P observed indicator scores, $\boldsymbol{\tau}$ is a $P \times 1$ vector of indicator intercepts, and $\boldsymbol{\Lambda}$ is a $P \times R$ matrix containing factor loadings of R common factors with means $\boldsymbol{\alpha}$ and covariance matrix $\boldsymbol{\Psi}$. Moreover, $\boldsymbol{t}_i$ is a $R \times 1$ vector of common-factor scores for individual i and $\boldsymbol{\varepsilon}_i$ is a $P \times 1$ vector of residual scores with variances of $\boldsymbol{\theta}$. The different levels of measurement invariance with respect to V ordered from least to most restrictive are

1 *Configural invariance*: implies equal factor structures across V,

2 *Metric invariance*: additionally implies equal factor loadings across V,

3 *Scalar invariance*: additionally implies equal indicator intercepts across V,

4 *Strict invariance*: additionally implies equal residual variances across V.

The model for the observed indicator scores and the measurement-invariance conditions easily generalize to cases with a unidimensional factor model.

One of the traditional methods to evaluate measurement invariance with respect to a categorical background variable (e.g., group membership) is MGCFA (Vandenberg & Lance, 2000). In MGCFA, the data are divided into two independent groups and a CFA model as shown in Equation 5.2 is estimated for each group. Each group thus has its own set of model parameters, which can be denoted as $\boldsymbol{\tau}^{(1)}$, $\boldsymbol{\Lambda}^{(1)}$, and $\boldsymbol{\theta}^{(1)}$ for Group 1

and $\boldsymbol{\tau}^{(2)}$, $\boldsymbol{\Lambda}^{(2)}$, and $\boldsymbol{\theta}^{(2)}$ for Group 2. Measurement invariance can then be assessed by comparing the fit of models with and without increasingly restrictive equality constraints on the parameters across the grouping variable. Because this method relies on splitting the data into two groups, it is best suited to a single categorical background variable. Alternative methods for assessing measurement invariance have been proposed that allow for including multiple categorical and continuous background variables simultaneously. Among these alternatives are restricted factor analysis (RFA; Oort, 1992), multiple indicator multiple cause (MIMIC; Jöreskog & Goldberger, 1975), and MNLFA (Bauer & Hussong, 2009) models. The difference between these methods and MGCFA is that the data are aggregated over the groups. We therefore refer to these methods as single-group methods (Kolbe et al., 2021), the most flexible of which is MNLFA, described below. For details about the differences between the single-group methods see Bauer (2017) and Kolbe et al. (2021).

5.2.2   Moderated Nonlinear Factor Analysis

In the MNLFA approach, a CFA model is estimated in which background variables are included as moderator variables. Figure 5.1 illustrates an example of a multidimensional MNLFA model containing two common factors, T_1 and T_2, measured by five indicators each. The idea of parameter moderation is demonstrated conceptually with the arrow pointing from V to the measurement model. In the following paragraphs, we consider a single background variable for ease of understanding, but MNLFA also allows for a set of multiple background variables.

More formally, the MNLFA model for continuous indicators X can be expressed as

$$\boldsymbol{x}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i \boldsymbol{t}_i + \boldsymbol{\varepsilon}_i, \tag{5.3}$$

where all parameters and notation remain defined as before in Equation 5.2, with the exception that all model parameters now have subscripts i. These subscripts are of special interest, because they indicate that the values of the parameters can differ over individuals as a function of the background variable V. In fact, all parameters of the MNLFA model can be allowed to differ across values of any observed variable, including residual (co)variances $\boldsymbol{\Theta}_i$, common-factor means $\boldsymbol{\alpha}_i$, and common-factor (co)variances $\boldsymbol{\Psi}_i$. Note that in this tutorial, we only consider linear relationships between the parameters and background variables, but other functional forms (e.g., quadratic, interaction) can be modeled as well (Bauer, 2017; Bauer et al., 2020).

The MNLFA model presumes configural invariance, but other levels of measurement invariance can be evaluated by testing whether V moderates any intercepts, factor loadings, or residual variances. More specifically, to accommodate any violations of scalar invariance, the vector of indicator intercepts $\boldsymbol{\tau}_i$ can be modeled as

$$\boldsymbol{\tau}_i = \boldsymbol{\tau}_0 + \boldsymbol{b} v_i, \tag{5.4}$$
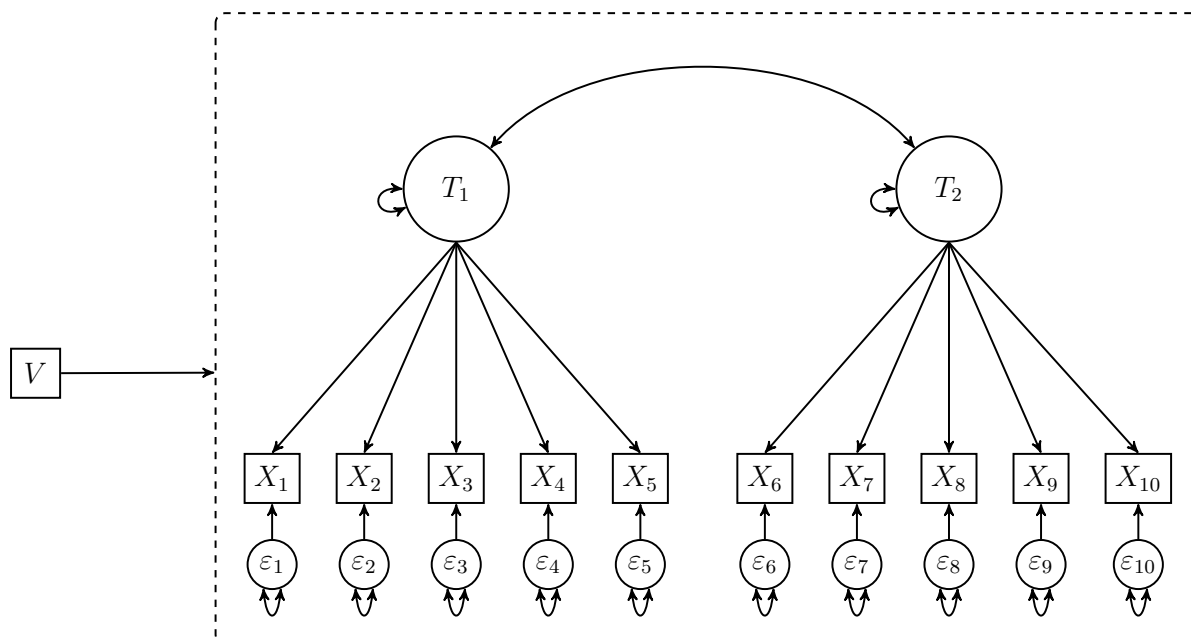
Figure 5.1: An example MNLFA model for assessing measurement invariance. The variable V may have an effect on all parameters in the model that is represented in the dashed border.

where $\boldsymbol{\tau}_0$ is a $P \times 1$ vector of baseline intercepts, $\boldsymbol{b}$ is a $P \times 1$ vector of linear effects of the background variable on the intercepts, and v_i is individual i's score on the background variable. A nonzero element in $\boldsymbol{b}$ reflects a linear change in the intercept associated with v_i, indicating a violation of scalar invariance (i.e., uniform DIF).

Similarly, to accommodate violations of metric invariance, each column of the matrix of factor loadings $\boldsymbol{\Lambda}_i$ can be modeled as a function of v_i

$$\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_0 + \boldsymbol{C}v_i, \tag{5.5}$$

where $\boldsymbol{\Lambda}_0$ is a $P \times R$ matrix of baseline factor loadings and $\boldsymbol{C}$ is a $P \times R$ matrix of linear effects of V on the factor loadings. A nonzero element in $\boldsymbol{C}$ reflects a linear change in the factor loadings associated with v_i, indicating a violation of metric invariance (i.e., nonuniform DIF).

The same idea can be adopted to accommodate violations of strict invariance. In order to prevent negative values of the residual variances, the indicators' residual variances $\boldsymbol{\theta}_i$ can be expressed as exponential functions of v_i

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 \exp(\boldsymbol{d}v_i), \tag{5.6}$$

where $\boldsymbol{\theta}_0$ is a $P \times 1$ vector of baseline residual variances and $\boldsymbol{d}$ is a $P \times 1$ vector containing the effects of v_i on the residual variances. A nonzero element in $\boldsymbol{d}$ indicates a violation of strict invariance.

In addition to measurement parameters, common-factor means, variances, and covari-

ances may also be moderated by the background variable V. The vector of means of the common factors $\boldsymbol{\alpha}_i$ may be written as a linear function of v_i

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_0 + \boldsymbol{g}v_i, \qquad (5.7)$$

where $\boldsymbol{\alpha}_0$ is a $R \times 1$ vector containing the baseline common-factor means and $\boldsymbol{g}$ is a $R \times 1$ vector containing the linear effects of v_i on the common-factor means. Moreover, the variances of the common factors may be moderated by the background variable. For example, the common-factor variance $\psi_{(T_1T_1)i}$ of common factor T_1 may be written as a log-linear function of v_i

$$\psi_{(T_1T_1)i} = \psi_{(T_1T_1)0}\exp(h_{(T_1T_1)}v_i). \qquad (5.8)$$

Here, $\psi_{(T_1T_1)0}$ is the baseline common-factor variance and $h_{(T_1T_1)}$ reflects the direct effect of the background variable on the common-factor variance. A nonzero h indicates that the common-factor variance differs across different levels of the background variable (i.e., common-factor heteroskedasticity). Similar to the model for the residual variances, a log-linear function is considered for the common-factor variances in order to prevent obtaining negative values. The baseline common-factor means and variances are commonly fixed at zero and one, respectively, in order to establish the origin and scale of the common factors.

Bauer (2017) suggested to model the covariance between common factors indirectly through a Fisher's z transformation to impose bounds of -1 and 1 on the corresponding correlation. So, in order to obtain the effect of the background variable on the covariance between common factors T_1 and T_2, the Fisher-transformed correlation between the factors $\zeta_{(T_1T_2)i}$ can be written as a linear function of v_i

$$\zeta_{(T_1T_2)i} = \zeta_{(T_1T_2)0} + h_{(T_1T_2)}v_i, \qquad (5.9)$$

where $\zeta_{(T_1T_2)0}$ is the baseline Fisher-transformed correlation between T_1 and T_2 and $h_{(T_1T_2)}$ reflects the direct effect of the background variable on the Fisher-transformed correlation. This Fisher-transformation can be inverted in order to obtain the correlation $\rho_{(T_1T_2)i}$ with bounds of -1 and 1

$$\rho_{(T_1T_2)i} = \frac{\exp(2\zeta_{(T_1T_2)i}) - 1}{\exp(2\zeta_{(T_1T_2)i}) + 1} \qquad (5.10)$$

and transformed to the covariance $\psi_{(T_1T_2)i}$ between the two common factors

$$\psi_{(T_1T_2)i} = \psi_{(T_1T_1)i}^{1/2}\rho_{(T_1T_2)i}\psi_{(T_2T_2)i}^{1/2}. \qquad (5.11)$$

The same approach could be applied to covariances between the residual factors of the indicators if present in the model.

Whereas RFA and MIMIC can only model specific parameters as functions of the background variable (i.e., common-factor means, indicators' intercepts, and factor loadings), MNLFA also allows for unique- and common-factor (co)variances to vary as functions

of the background variable. The MNLFA approach can thus be conceptualized as an extended RFA or MIMIC model in which (co)variances are not necessarily homoskedastic across different levels of the background variable. This makes MNLFA as flexible as MGCFA when the background variable is dichotomous, yet more flexible because MNLFA also allows for multiple categorical and continuous background variables to be included in a single model. For more details on the MNLFA model (e.g., parameter equations for situations with multiple background variables), see Bauer (2017).

5.3 Tutorial

The empirical data that we used for this tutorial were gathered by Denollet et al. (2013). The data contain observed scores on the DS14 (Denollet, 2005) in a sample of 541 patients with coronary artery disease. The DS14 is a widely used instrument for the assessment of the Type D personality and consists of 14 items, of which seven measure negative affectivity and seven measure social inhibition. All items have five ordered response categories ($0 = $ *false*, $1 = $ *rather false*, $2 = $ *neutral*, $3 = $ *rather true*, $4 = $ *true*). In addition to the observed scores on the DS14, the data also contain the dichotomous variable gender and the continuous variable age measured in years. These two variables were used as background variables in this tutorial.

The MNLFA model considered in this tutorial is shown in Figure 5.2. The model includes two common factors, social inhibition (SI) and negative affectivity (NA), each measured by seven indicators. The two common factors are allowed to covary and their means and variances are fixed at zero and one, respectively, in order for the model to be identified. The background variables gender and age are included in the model as moderators, and depending on the level of measurement invariance may or may not moderate the indicators' intercepts and factor loadings. In almost all MNLFA models in this tutorial, the common-factor means, common-factor variances, common-factor covariance, and indicators' residual variances are allowed to vary as a function of gender and age[2]. Such models are also referred to as heteroskedastic MNLFA models (see Kolbe et al., 2021).

In the next section of the chapter, we demonstrate how to evaluate whether the DS14 is measurement invariant with respect to gender and age using MNLFA. We provide a step-by-step tutorial for assessing full measurement invariance (i.e., assessing measurement invariance of all indicators simultaneously), selecting anchor indicators, and evaluating partial invariance (i.e., testing each indicator separately for measurement invariance). We focus on how to detect violations of scalar and metric invariance (i.e., uniform and nonuniform DIF, respectively). Such violations are commonly examined with separate omnibus tests of scalar and metric invariance, but can also be assessed simultaneously rather than separately with a single omnibus test (Putnick & Bornstein, 2016; B. Muthén & Asparouhov, 2002; Stark et al., 2006). In this single omnibus test, the violations

---

[2]An exception is the configural model, which requires additional identification constraints. This will be explained in Step 3a of the tutorial.
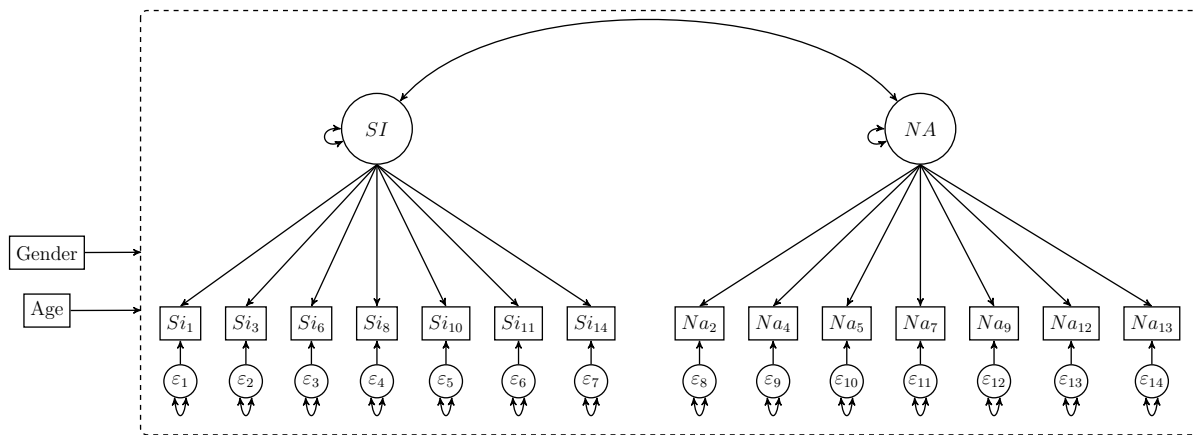
Figure 5.2: The MNLFA model for the DS14 dataset.

of scalar and metric invariance are examined simultaneously by comparing the fit of a configural model (that allows for uniform and nonuniform DIF in all indicators) to the fit of a scalar model (that does not allow for uniform or nonuniform DIF in any indicators). Full R scripts for replicating the results of this tutorial can also be found on our Open Science Framework project `https://osf.io/6cyxt/`.

Before we present all of the steps of the tutorial, we would like to make two important comments. First, although the measurement level of the DS14 items is ordinal, we treat the items as continuous in the present tutorial because our aim is merely to illustrate MNLFA and not to draw substantive conclusions. Please note that if MNLFA were to be used for drawing substantive conclusions with respect to ordinal data, an ordinal MNLFA model may be more suitable (see section **Extensions**). For more details on treating ordinal items as continuous, see Rhemtulla et al. (2012) and Robitzsch (2020b). Rhemtulla et al. (2012) discussed conditions under which 5-point scales might be treated as continuous, and Robitzsch (2020b) elucidated why it might always be defensible to do so. Second, there is no straightforward way of assessing configural invariance (nor overall data–model fit in general) with MNLFA yet, because MNLFA does not allow for distinct factor structures with respect to the background variable(s). We will therefore assume in this tutorial that the configural-invariance condition holds. In practice, it is important to evaluate configural invariance prior to the assessment of metric, scalar, and strict invariance, because the assessment of other levels of measurement invariance may lead to false conclusions if configural invariance does not hold (Jorgensen, 2017).

5.3.1 Step 1: Install and Load OpenMx

This tutorial illustrates how measurement invariance can be examined with MNLFA using `OpenMx` (version 2.19.5; Neale et al., 2016). This package can be used for matrix algebra optimization, SEM, and other statistical estimation methods. The user needs to install the package on the computer once, then load it into the workspace each time a new R session is started:

```
> install.packages("OpenMx")
> library(OpenMx)
```

Any dependencies are automatically installed when running the syntax above. Note that R version 3.5 or higher is required[3] for the latest version of this package to work. The core function of the `OpenMx` package necessary to create an MNLFA model is the `mxModel()` function. This function builds an object for a statistical model containing (among other information) the data, matrices, algebraic expressions, fit functions, and expectations for the model.

5.3.2 Step 2: Load and Prepare Data

The data gathered by Denollet et al. (2013) is available in the `mokken` package (Van der Ark, 2007). This R package can be installed and loaded as follows:

```
> install.packages("mokken")
> library(mokken)
```

In order to load the DS14 data and convert it to a data frame, the user can run the following lines of R code:

```
> data("DS14", package="mokken")
> DS14 <- data.frame(DS14)
```

The dataset `DS14` is now loaded into the user's workspace. The `head()` function can be used to inspect the first six rows of the data. The data contain gender (`Male`), age (`Age`), and item scores on the DS14 questionnaire (`Si1.`, `Na2`, `Si3.`, `Na4`, `Na5`, `Si6`, `Na7`, `Si8`, `Na9`, `Si10`, `Si11`, `Na12`, `Na13`, `Si14`).

For ease of interpretation, the two negatively worded items (`Si1.` and `Si3.`) can be re-coded prior to the measurement invariance assessment:

```
> DS14$Si1 <- 4 - DS14$Si1.
> DS14$Si3 <- 4 - DS14$Si3.
```

After running the above syntax, the `DS14` data frame now also contains the re-coded versions of these two items, named `Si1` and `Si3`. Similar to the other SI items, a higher score on these re-coded items now indicates a higher level of social inhibition. In addition to the item scores, the data contain scores on the variables gender (`Male`) and age (`Age`). A score of `Male = 1` represents a male patient and `Male = 0` a female patient. The `Age` variable is measured in years and can be standardized in order for an easier interpretation of its effects on the model parameters:

```
> DS14$Age <- (DS14$Age - mean(DS14$Age))/sd(DS14$Age)
```

For convenience, the item scores are re-ordered such that the first seven items reflect social inhibition and the second seven items reflect negative affectivity:

```
> DS14 <- DS14[,c("Male","Age","Si1","Si3","Si6","Si8","Si10","Si11","Si14",
>                  "Na2","Na4","Na5","Na7","Na9","Na12", "Na13")]
```

---

[3]This may have changed since the time of this writing. Check `https://CRAN.R-project.org/package=OpenMx` for the current requirements.

After the dataset has been prepared for the measurement invariance assessment, the user should convert the data to an `MxData` object:

```
> mxdata1 <- mxData(observed=DS14, type="raw")
```

The `mxData()` function constructs an object with additional information allowing it to be processed in the `mxModel()` function. By specifying `observed=DS14` and `type="raw"`, the function reads in the raw data stored in `DS14`. Alternatively, summary statistics could be analyzed.

Finally, the user can save the number and names of the observed variables serving as indicators of the factors in the MNLFA model:

```
> manVars <- colnames(DS14[,-c(1,2)])
> nv <- length(manVars)
```

The names of the indicators stored in `manVars` are required for one of the arguments of the `mxModel()` function, which will be shown later in this section.

5.3.3   Step 3: Assess Full Measurement Invariance

After preparing the data, the user can start with performing an omnibus test of full measurement invariance with respect to the background variables. That is, the user can test the H_0 that none of the indicators function differently with respect to `Male` and `Age`. Full measurement invariance can be assessed with MNLFA on metric, scalar, or strict levels. In this step of the tutorial, we will focus only on simultaneously assessing full scalar-and-metric invariance. In the omnibus test of scalar-and-metric invariance, the fit of an unconstrained configural model is compared to the fit of a constrained scalar model. In the configural model, the direct effects of the background variables on the indicators' intercepts and factor loadings are all freely estimated, whereas in the scalar model these effects are all fixed to zero. If the scalar model fits the data significantly worse than the configural model at a chosen α level, the H_0 of full scalar-and-metric invariance is rejected and follow-up tests can be performed to evaluate which specific indicators exhibit (non)uniform DIF with respect to the background variables (Steps 4 and 5).

Step 3a: Specify and Fit the Configural Model

In order to fit this model to the empirical data with `OpenMx`, we put all model parameters into matrices using the `mxMatrix()` function. Most often, the following six arguments will be specified for each matrix:

`type`: requires a character string indicating the matrix type. In this tutorial, we use `"Diag"`, `"Full"`, and, `"Symm"` matrices.

`nrow`: refers to the number of rows of the matrix.

`ncol`: refers to the number of columns of the matrix.

free: indicates which elements of the matrix can be freely estimated (`TRUE` or `T`) or are fixed parameters (`FALSE` or `F`).

values: reflects the values of the elements in the matrix. If an element is freely estimated, it reflects the starting value. If an element is not freely estimated, it reflects the fixed value.

name: refers to the user-specified name of the matrix which is used within `OpenMx` when performing an operation on this matrix.

The syntax below creates three `MxMatrix` objects for the indicator intercepts of the configural model:

```
> matT0 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=TRUE,
>                    values=1,
>                    name="matT0")
> matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=TRUE,
>                    values=0,
>                    name="matB1",
> matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=TRUE,
>                    values=0,
>                    name="matB2"
```

The matrix `matT0` is a full matrix containing the baseline intercepts $\boldsymbol{\tau}_0$. All baseline intercepts are freely estimated with starting values of one by setting `free=TRUE` and `values=1`. Matrix `matB1` and `matB2` are full matrices containing the direct effects of the background variables `Male` and `Age`, respectively, on the intercepts. These direct effects reflect uniform DIF, represented by b. In the configural model, the effects of `Male` and `Age` on the intercepts are freely estimated with starting values of zero by setting `free=TRUE` and `values=0`. By giving the matrices names using the `name` argument, we can refer to these matrices in upcoming syntax.

Similar lines of R code can be used for creating the matrices of factor loadings:

```
> matL0 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                    free=c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7)),
>                    values=c(rep(c(1,0),7), rep(c(0,1),7)),
>                    byrow=TRUE,
>                    name="matL0")
> matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                    free=c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7)),
>                    values=0,
>                    byrow=TRUE,
>                    name="matC1")
> matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                    free=c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7)),
>                    values=0,
>                    byrow=TRUE,
>                    name="matC2")
```

Here, `matL0` is a full matrix containing the baseline factor loadings $\boldsymbol{\Lambda}_0$. The first column of the matrix contains the factor loadings of the social inhibition factor, and the second column contains the factor loadings of the negative affectivity factor. By setting

`free=c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7))`, `values=c(rep(c(1,0),7), rep(c(0,1),7))`, and `byrow=TRUE`, the factor loadings of the social inhibition factor on the first seven items and the factor loadings of the negative affectivity factor on the second seven items are freely estimated (with starting values = 1). For the other factor-indicator combinations, the factor loadings are fixed at zero. Matrices `matC1` and `matC2` are full matrices containing the direct effects of `Male` and `Age`, respectively, on the factor loadings (i.e., nonuniform DIF represented by c). These effects are freely estimated in the configural model with starting values of zero by setting `free=c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7))`, `values=0`, and `byrow=TRUE`.

The matrices for the residual variances of the indicators can then be specified as symmetric (`type="Symm"`) if there are any nonzero residual covariances, or more simply diagonal (`type="Diag"`) if only residual variances are nonzero (which is the case here):

```
> matE0 <- mxMatrix(type="Diag", nrow=nv, ncol=nv,
>                   free=TRUE,
>                   values=1,
>                   name="matE0")
> matD1 <- mxMatrix(type="Diag", nrow=nv, ncol=nv,
>                   free=TRUE,
>                   values=0,
>                   name="matD1")
> matD2 <- mxMatrix(type="Diag", nrow=nv, ncol=nv,
>                   free=TRUE,
>                   values=0,
>                   name="matD2")
```

where `matE0` is a diagonal matrix containing the baseline residual variances $\boldsymbol{\theta}_0$, `matD1` is a diagonal matrix containing the effects of `Male` on the residual variances, and `matD2` is a diagonal matrix containing the effects of `Age` on the residual variances (also represented as d). These model parameters are all freely estimated with the `free=TRUE` argument.

After specifying the matrices of measurement parameters, matrices can be specified for the common-factor variances and correlation. These matrices can be specified as follows:

```
> matP0 <- mxMatrix(type="Symm", nrow=2, ncol=2,
>                   free=c(FALSE,TRUE,TRUE,FALSE),
>                   values=c(1,0,0,1),
>                   name="matP0")
> matH1 <- mxMatrix(type="Symm", nrow=2, ncol=2,
>                   free=c(FALSE,TRUE,TRUE,FALSE),
>                   values=0,
>                   name="matH1")
> matH2 <- mxMatrix(type="Symm", nrow=2, ncol=2,
>                   free=c(FALSE,TRUE,TRUE,FALSE),
>                   values=0,
>                   name="matH2")
```

where `matP0` is a symmetric matrix containing the baseline common-factor variances and the baseline correlation[4] between the two common factors. We freely estimate the baseline common-factor correlation and fix the baseline common-factor variances to unity in order

---

[4]Combining the common-factor variances and correlations in a single matrix may seem odd, but allows us to specify different moderation functions for variance parameters versus correlations (see Bauer, 2017). This is shown in a later paragraph of Step 3a.

for the scales of the common factors to be identified. Matrices `matH1` and `matH2` contain the direct effects of `Male` and `Age`, respectively, on the common-factor variances and correlation (also denoted by h). In the configural model we set `free=c(FALSE,TRUE,TRUE,FALSE)` and `values=0` in `matH1` and `matH2`, so the direct effects of the background variables on the common-factor correlation are freely estimated, but the direct effects on the common-factor variances are fixed to zero.

The matrices required for the common-factor means can be specified with the following R syntax:

```
> matA0 <- mxMatrix(type="Full", nrow=2, ncol=1,
>                   free=FALSE,
>                   values=0,
>                   name="matA0")
> matG1 <- mxMatrix(type="Full", nrow=2, ncol=1,
>                   free=FALSE,
>                   values=0,
>                   name="matG1")
> matG2 <- mxMatrix(type="Full", nrow=2, ncol=1,
>                   free=FALSE,
>                   values=0,
>                   name="matG2")
```

Here, `matA0` is a matrix containing the baseline common-factor means, which are fixed at zero for the origins of the common factors to be identified. The `matG1` and `matG2` matrices contain the direct effects of `Male` and `Age`, respectively, on the common-factor means (also represented by g). These direct effects are fixed at zero in the configural model. So, in addition to fixing the baseline common-factor variances and means to one and zero, respectively, the direct effects of the background variables on the common-factor variances and means are also fixed at zero in the configural model. These additional identification constraints are necessary because the configural model includes all possible direct effects of the background variables on the indicators' intercepts, factor loadings, and residual variances (analogous to measurement parameters differing across groups in a configural MGCFA model).

Now that all the matrices with baseline parameters and direct effects of the background variables on the parameters have been created, the user can specify the matrices required for the matrix algebra in the next paragraphs. First, to allow for moderating effects, the background variables are modeled as definition variables with the following matrices:

```
> matV1 <- mxMatrix(type="Full", nrow=1, ncol=1,
>                   free=FALSE,
>                   labels="data.Male",
>                   name="Male")
> matV2 <- mxMatrix(type="Full", nrow=1, ncol=1,
>                   free=FALSE,
>                   labels="data.Age",
>                   name="Age")
```

The matrices `matV1` and `matV2` contain the observed scores on the background variables `Male` and `Age`, respectively. The observed scores on the background variables stored in the `mxdata1` dataset are referred to in the matrix label as `"data.Male"` and `"data.Age"`. Modeling the background variables as definition variables allows us to let model parameters differ

across different levels of `Male` and `Age`.

Then, the matrices for all parameters predicted by `Male` and `Age` are created using the `mxAlgebra()` function. The `mxAlgebra()` function can be used to define a matrix of model parameters as a function of background variables. The first argument of this function, `expression`, should be used for specifying an R expression of one or more `MxMatrix` objects. Most R operators like `+`, `*`, and `%*%`, an general R functions like `mean()`, `log()`, and `exp()` are supported in this argument of the `mxAlgebra()` function. A name for the defined matrix can be assigned with the `name` argument. The matrices of intercepts, factor loadings, and residual variances (respectively) can be specified as:

```
> matT <- mxAlgebra(expression = matT0 + matB1*Male + matB2*Age,
>                    name="matT")
> matL <- mxAlgebra(expression = matL0 + matC1*Male + matC2*Age,
>                    name="matL")
> matE <- mxAlgebra(expression = matE0*exp(matD1*Male + matD2*Age),
>                    name="matE")
```

The object `matT` contains a linear moderation function for the indicator intercepts $\boldsymbol{\tau}$ of patient i (as in Equation 5.4):

$$\boldsymbol{\tau}_i = \boldsymbol{\tau}_0 + \boldsymbol{b}_1 \times Male_i + \boldsymbol{b}_2 \times Age_i, \tag{5.12}$$

and `matL` contains a linear moderation function for the factor loadings $\boldsymbol{\Lambda}$ of patient i (as in Equation 5.5):

$$\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_0 + \boldsymbol{C}_1 \times Male_i + \boldsymbol{C}_2 \times Age_i. \tag{5.13}$$

The object `matE` contains a log-linear function for the residual variances $\boldsymbol{\theta}$ of patient i (as in Equation 5.6):

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_0 \times \exp(\boldsymbol{d}_1 \times Male_i + \boldsymbol{d}_2 \times Age_i). \tag{5.14}$$

Similarly, the matrix of common-factor means can be created with the `mxAlgebra()` function:

```
> matA <- mxAlgebra(expression = matA0 + matG1*Male + matG2*Age,
>                    name="matA")
```

where `matA` contains the common-factor means $\boldsymbol{\alpha}$ being modeled as a linear function of the background variables:

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_0 + \boldsymbol{g}_1 \times Male_i + \boldsymbol{g}_2 \times Age_i. \tag{5.15}$$

Finally, the covariance matrix for the common factors requires different moderation functions for variances and covariance(s). First, we model the common-factor variances as a log-linear function of `Male` and `Age`:

```
> matVar <- mxAlgebra(expression = matP0*exp(matH1*Male + matH2*Age),
>                      name="matVar")
```

The object `matVar` thus contains a matrix with the common-factor variances on the diag-

onal. For example, the variance of the social inhibition factor is modeled as

$$\psi_{(SISI)i} = \psi_{(SISI)0} \times \exp(h_{(SISI)1} \times Male_i + h_{(SISI)2} \times Age_i). \qquad (5.16)$$

After we have specified the common-factor variances, we can obtain the common-factor covariance via the common-factor correlation. To respect a correlation's natural bounds between -1 and 1, we apply a Fisher's z transformation to the correlation, making it a linear function of the background variables (see Equation 5.10):

```
> matR <- mxAlgebra(expression=(exp(2*(matP0 + matH1*Male + matH2*Age)) - 1)/
>                             (exp(2*(matP0 + matH1*Male + matH2*Age)) + 1),
>                 name="matR")
```

The object `matR` includes a matrix with the common-factor correlation bound between -1 and 1 on the off-diagonal. Before converting factor correlations to covariances, the user must first make a 2×2 identity matrix (`matIa`), as well as a 2×2 matrix with zeros on the diagonal and ones on the off-diagonal (`matIb`).

```
> matIa <- mxMatrix(type="Diag", nrow=2, ncol=2,
>                   free=FALSE,
>                   values=1,
>                   name="matIa")
> matIb <- mxMatrix(type="Full", nrow=2, ncol=2,
>                   free=FALSE,
>                   values=c(0,1,1,0),
>                   name="matIb")
```

We need these matrices to set diagonal and off-diagonal elements of upcoming matrices to zero. Now, the correlation can be converted to a covariance as follows:

```
> matCov <- mxAlgebra(expression=(matIa*sqrt(matVar)) %*% matR %*%
>                   (matIa*sqrt(matVar)), name="matCov")
```

Here, we take the square root of the matrix `matVar` to obtain SDs on the diagonal, then premultiply this matrix by `matIa` to set all off-diagonal elements to zero. The `matCov` matrix contains the common-factor covariance on the off-diagonal.

Now we can add `matIa*matVar` (i.e., a matrix with common-factor variances on the diagonal and zeros on the off-diagonal) to `matIb*matCov` (i.e., a matrix with zeros on the diagonal and the common-factor covariance on the off-diagonal) to obtain the covariance matrix for the common factors:

```
> matP <- mxAlgebra(expression = matIa*matVar + matIb*matCov,
>                 name="matP")
```

With all of the matrices specified above (i.e., `matT`, `matL`, `matE`, `matA`, and `matP`), the model-implied moments (means and covariance matrix) of the configural model can be specified, again by using the `mxAlgebra()` function:

```
> matM <- mxAlgebra(expression = matT + t(matL)*matA,
>                 name="matM")
> matC <- mxAlgebra(expression = matL %*% matP %*% t(matL) + matE,
>                 name="matC")
```

The `matM` matrix contains the model-implied means, and the `matC` matrix contains the model-implied variances and covariances of the indicators.

In order to fit the configural model with the specified model-implied matrices, the user needs to specify the model expectation and fit function:

```
> expF <- mxExpectationNormal(covariance="matC",
>                             means="matM",
>                             dimnames=manVars)
> fitF <- mxFitFunctionML()
```

The expectation function stored in `expF` defines the way in which the model expectations are calculated. The `mxExpectationNormal()` function uses the algebra defined in `matC` and `matM` to obtain the model-implied variances, covariances, and means of the indicators under multivariate normality. The `dimnames` argument of the function takes the character vector `manVars` containing the names of the indicators. The `mxFitFunctionML()` function stored in `fitF` is used to indicate that the free parameters of the configural model should be estimated using full-information maximum likelihood. Alternatively, a user-defined fit function can be treated as an `mxFitFunction` by using the `mxFitFunctionR()` function.

All of the separate objects for each part of the configural model can now be combined using the `mxModel()` function. The `model` argument of this function can be used to specify a name for the new model. The following arguments are a number of `MxMatrix` and `MxAlgebra` objects, as well as the expectation function, fit function, and `MxData` objects. All of these objects can be added to the model as follows:

```
> modConfig <- mxModel(model="Configural",
>                      matT, matT0, matB1, matB2,
>                      matL, matL0, matC1, matC2,
>                      matE, matE0, matD1, matD2,
>                      matP, matP0, matH1, matH2,
>                      matA, matA0, matG1, matG2,
>                      matIa, matIb, matV1, matV2,
>                      matVar, matR, matCov, matM, matC,
>                      expF, fitF, mxdata1)
```

The object `modConfig` now includes the data, model matrices, expectation function, and fit function. These objects are all the required elements to optimize the free parameters in the model. The model can be fitted to the data using the `mxRun()` function. This function sends the `MxModel` object specified in the first argument to the optimizer and returns the optimized model. Additional information on the parameters estimates can be requested by including arguments like `intervals = TRUE` for confidence intervals. The configural model can be fitted to the DS14 data as follows:

```
> fitConfig <- mxRun(modConfig)
```

The output can be printed using the `summary()` function. The summary output of the model contains the estimates of all free parameters, their standard errors, and model statistics including the number of parameters and goodness-of-fit reported in units of -2 times the log-likelihood (`-2lnL`). In Step 3c, this fit statistic will be compared to the fit of the scalar model (Step 3b).

Step 3b: Specify and Fit the Scalar Model

In this step, we show how to specify and fit the scalar model in which all direct effects of `Male` and `Age` on the indicators' intercepts and factor loadings are fixed at zero. The user should first re-specify matrices `matB1`, `matB2`, `matC1`, and `matC2`:

```
> matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                   free=FALSE,
>                   values=0,
>                   name="matB1")
> matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                   free=FALSE,
>                   values=0,
>                   name="matB2")
> matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                   free=FALSE,
>                   values=0,
>                   name="matC1")
> matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                   free=FALSE,
>                   values=0,
>                   name="matC2")
```

These new matrices for the direct effects on the intercepts and factor loadings indicate that none of the direct effects of the background variables on the indicator's intercept and factor loading should be freely estimated by using the `free=FALSE` argument in each of these matrices.

Next, we release the additional identification constraints that were necessary for the configural model in Step 3a. That is, in the scalar model we freely estimate the direct effects of `Male` and `Age` on the common-factor means and variances by re-specifying matrices `matH1`, `matH2`, `matG1`, and `matG2`:

```
> matH1 <- mxMatrix(type="Symm", nrow=2, ncol=2,
>                   free=TRUE,
>                   values=0,
>                   name="matH1")
> matH2 <- mxMatrix(type="Symm", nrow=2, ncol=2,
>                   free=TRUE,
>                   values=0,
>                   name="matH2")
> matG1 <- mxMatrix(type="Full", nrow=2, ncol=1,
>                   free=TRUE,
>                   values=0,
>                   name="matG1")
> matG2 <- mxMatrix(type="Full", nrow=2, ncol=1,
>                   free=TRUE,
>                   values=0,
>                   name="matG2")
```

These matrices now indicate that the common-factor variances and means are allowed to differ across all values of `Male` and `Age`. All other elements of the scalar model are similar to the configural model specified in Step 3a. So, after re-specifying objects `matB1`, `matB2`, `matC1`, `matC2`, `matH1`, `matH2`, `matG1`, and `matG2` we can combine the elements required to run the scalar model and fit the model to the data:

```
> modScalar <- mxModel(model="Scalar",
>                      matT, matT0, matB1, matB2,
```

```
>                           matL, matL0, matC1, matC2,
>                           matE, matE0, matD1, matD2,
>                           matP, matP0, matH1, matH2,
>                           matA, matA0, matG1, matG2,
>                           matIa, matIb, matV1, matV2,
>                           matVar, matR, matCov, matM, matC,
>                           expF, fitF, mxdata1)
> fitScalar <- mxRun(modScalar)
```

Again, the `summary()` function can be used to obtain the model fit and parameter estimates.

Step 3c: Conduct Likelihood-Ratio Test

Now that both the configural and scalar model have been fitted to the data, a likelihood-ratio test (LRT) can be performed using the `mxCompare()` function:

```
> miTest <- mxCompare(fitConfig, fitScalar)
> miTest
        base comparison  ep minus2LL   df     AIC   diffLL diffdf          p
1 Configural       <NA> 129 20762.42 7435 21020.42       NA     NA         NA
2 Configural      Scalar  81 20874.94 7483 21036.94 112.5204     48 4.253424e-07
```

Using $\alpha = .05$ as significance level, the output shows that the constraints on the intercepts and factor loadings significantly deteriorate model fit, $\Delta\chi^2(48) = 112.52, p < .001$. This indicates that the H_0 of full scalar-and-metric invariance can be rejected. In the following steps, we will illustrate how follow-up tests can be performed to evaluate which indicators function differently with respect to the background variables.

5.3.4   Step 4: Select Anchor Indicators

In the previous step, we have rejected the H_0 of full scalar-and-metric invariance with respect to `Male` and `Age`. Partial invariance can be established by detecting which specific indicators exhibit (non)uniform DIF. Each indicator can be tested individually for DIF while holding a subset of other indicators invariant across the background variables. These latter indicators are also called anchor indicators and are used to link the metric of the common factors across the background variables. When anchor indicators are not known a priori, they can be explicitly selected using an anchor-selection strategy. In this step of the tutorial, we show how to apply the rank-based strategy (Woods, 2009) for the selection of anchor indicators. This is an easily implemented selection strategy in which a limited number of indicators that show the weakest evidence of DIF are selected as anchor indicators. More complicated empirical methods for selecting anchors can slightly improve accuracy (Kopf et al., 2015b,a), and regularized estimation can avoid the need for anchors altogether (Bauer et al., 2020). The potential danger of selecting anchor indicators is that one or more indicators with DIF may be selected as anchors, which can cause problems such as biased parameter estimates and an overestimation of the amount of DIF (W.-C. Wang, 2004). However, the risk of bias can be minimal if positive and negative DIF are relatively balanced. When DIF is unbalanced, the risk of such a contamination of the subset of anchor indicators can be decreased by selecting a relatively small anchor

set. Accordingly, Woods' (2009) recommended to select approximately 20% of indicators to serve as anchor indicators. We follow this recommendation by selecting two out of seven indicators per common factor to serve as anchor indicators across both background variables[5]. One could also argue to select anchor indicators for each background variable separately, but to keep the following steps of the tutorial as concise as possible we demonstrate how to select the same anchor indicators for `Male` and `Age`.

Step 4a: Specify and Fit All-But-One Models

The first step of the rank-based strategy is to fit a less-constrained all-but-one model for each indicator. In an indicator's all-but-one model, only that indicator's intercept and factor loading are predicted by the background variables (i.e., all but one of the indicators are assumed to be scalar-and-metric invariant). For example, the all-but-one model for Indicator 1 (`si1`) includes freely estimated direct effects of `Male` and `Age` on the intercept and factor loading of Indicator 1, and the direct effects of the background variables on all other indicators' intercepts and factor loadings are fixed at zero. So, almost all of the matrices specified in Step 3a can be used for these models, except for matrices `matB1`, `matB2`, `matC1`, and `matC2`.

In order to efficiently execute this step, we specify and fit an all-but-one model for each indicator in a `for` loop. First, we create an empty list to which we can add each model's output:

```
> fitAbo <- list()
```

Next, we run the following `for` loop:

```
> for (i in 1:nv){
>     freeparT <- matrix(data=FALSE, nrow=1, ncol=nv)
>     freeparT[i] <- TRUE
>     freeparL <- matrix(data=FALSE, nrow=nv, ncol=2)
>     freeparL[i, ifelse(i < 8, yes=1, no=2)] <- TRUE
>     matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                       free=freeparT,
>                       values=0,
>                       name="matB1")
>     matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                       free=freeparT,
>                       values=0,
>                       name="matB2")
>     matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                       free=freeparL,
>                       values=0,
>                       byrow=TRUE,
>                       name="matC1")
>     matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
```

---

[5]Note that if only one anchor indicator per common factor is used for identification, the model would simply be a configural model, equivalent to the one specified in Step 3a. Such a model is incapable of distinguishing between differences in common-factor means and variances from differences in the reference-indicator's intercept and loading across the background variables. That is why at least two anchor indicators are required per common factor for differences in common-factor means and variances to be attributed to differences in the common-factor distribution, rather than being due to differences in a single indicator.

```
>                          free=freeparL,
>                          values=0,
>                          byrow=TRUE,
>                          name="matC2")
>      modAbo <- mxModel(model=paste0("All_but_", i),
>                        matT, matT0, matB1, matB2,
>                        matL, matL0, matC1, matC2,
>                        matE, matE0, matD1, matD2,
>                        matP, matP0, matH1, matH2,
>                        matA, matA0, matG1, matG2,
>                        matIa, matIb, matV1, matV2,
>                        matVar, matR, matCov, matM, matC,
>                        expF, fitF, mxdata1)
>      fitAbo[[i]] <- mxRun(modAbo)
> }
```

In this `for` loop, the syntax will run for each of the 14 indicators represented by `i`. First, we specify which elements in `matB1`, `matB2`, `matC1`, and `matC2` should be freely estimated by creating matrices with true and false entries, named `freeparT` and `freeparL`. These matrices are used for the `free` argument of the `mxMatrix()` function to indicate that the direct effects of the background variables on indicator `i`'s intercept and factor loading should be freely estimated. The effects of the background variables on all intercepts and factor loadings of indicators other than `i` are fixed to zero (implying no DIF). After re-specifying the matrices `matB1`, `matB2`, `matC1`, and `matC2`, all elements required to fit the all-but-one model of indicator `i` are combined using the `mxModel()` function. The model is then optimized using `mxRun()` and added to the `i`th component of the list `fitAbo`.

Step 4b: Conduct Likelihood-Ratio Tests and Select Anchors

After specifying and fitting the less-constrained all-but-one models, each one's fit can be compared to the fit of the constrained scalar-invariance model (`fitScalar`). A comparison of the fit of these models with an LRT indicates whether additionally fixing the current indicator's intercept and factor loading to be unaffected by the background variables leads to a significantly worse model fit. Note that these LRTs should not be trusted as tests of DIF because unmodeled DIF biases other parameters, inflating Type I error rates for DIF-free indicators; however, these tests can serve as a reliable empirical basis for selecting anchors (Woods, 2009; Kolbe & Jorgensen, 2019; Kopf et al., 2015a). Because `fitAbo` is a list of fitted models, each of them will be compared to `fitScalar`, and we store the LRT results in a readable table `anchorOut`:

```
> anchorTest <- mxCompare(fitAbo, fitScalar)
> anchorOut <- data.frame(Name = paste0("Indicator", 1:nv),
>                         X2   = anchorTest$diffLL[seq(2,28,2)],
>                         df   = anchorTest$diffdf[seq(2,28,2)],
>                         p    = anchorTest$p[seq(2,28,2)])
```

Differences in `-2lnL` values from the scalar-invariance and all-but-one models follow a χ^2 distribution with $df = 4$. The results stored in `piOut` are presented in Table 5.1.

Indicators 1–7 of social inhibition and 8–14 of negative affectivity can now be ranked in ascending order based on their LRT statistics:

Table 5.1: Likelihood-ratio tests for the purpose of selecting anchor indicators.

Indicator	Name	$\Delta\chi^2(4)$	p	Rank
1	Si1	16.36	.003	6
2	Si3	8.78	.067	5
3	Si6	5.44	.245	3
4	Si8	7.42	.115	4
5	Si10	4.30	.367	2
6	Si11	3.05	.550	1
7	Si14	16.72	.002	7
8	Na2	2.43	.658	2
9	Na4	1.15	.886	1
10	Na5	11.02	.026	5
11	Na7	5.36	.252	4
12	Na9	3.90	.420	3
13	Na12	22.54	.000	7
14	Na13	19.38	.001	6

Note. The rank score is based on the LRT statistic (i.e., $\Delta\chi^2(4)$) of each indicator per common factor. The indicators are ranked in ascending order, that is, a higher rank score indicates a smaller LRT statistic and thus a weaker evidence of DIF.

```
> anchorOut[order(anchorOut$X2[1:7]),]
> anchorOut[7+order(anchorOut$X2[8:14]),]
```

The smaller the test statistic, the weaker the evidence of DIF. So, for each common factor, the two indicators with the smallest test statistics are selected as anchor indicators:

```
> anchors1 <- c(5, 6)
> anchors2 <- c(8, 9)
```

In this dataset, Indicators 5 and 6 are selected as anchor indicators for social inhibition, and Indicators 8 and 9 are selected as anchor indicators for negative affectivity. The indicator indices are stored in `anchors1` and `anchors2` to use as anchors in Step 5, when we test all other studied indicators (i.e., Indicators 1, 2, 3, 4, 7, 10, 11, 12, 13, 14) for DIF.

5.3.5   Step 5: Assess Partial Invariance

Previously in Step 3, we found evidence against full scalar-and-metric invariance with respect to the background variables `Male` and `Age`. We can now perform follow-up tests in order to evaluate partial scalar-and-metric invariance. We will show how to use MNLFA to test the H_0 of invariance for each indicator by comparing the fit of a less-constrained anchors-only model to several more-constrained anchors-plus-one models. In the anchors-only model, the direct effects of the background variables on all studied indicators' intercepts and factor loadings are freely estimated to allow for (non)uniform DIF, so only

the anchor indicators have scalar- and metric-invariance constraints. For each studied indicator, its anchors-plus-one model additionally constrains that indicator to be invariant by removing the background variables' (non)uniform DIF estimates. Because background variables continue to affect remaining studied indicators, parameter estimates in these anchors-plus-one models are not biased by DIF (unless selected anchors have DIF, which is a small risk; Woods, 2009; Kolbe & Jorgensen, 2019; Kopf et al., 2015b).

Step 5a: Specify and Fit the Anchors-Only Model

We first create an object containing the studied indicators, by removing anchors:

```
> testIn <- c(1:nv)[ -c(anchors1, anchors2) ]
```

Then, we create two matrices that indicate which direct effects of the background variables are freely estimated in the anchors-only model:

```
> freeparT <- matrix(TRUE, nrow=1, ncol=14)
> freeparT[1, c(anchors1, anchors2)] <- FALSE
> freeparL <- matrix(c(rep(c(TRUE,FALSE),7), rep(c(FALSE,TRUE),7)),
>                     nrow=nv, ncol=2, byrow=TRUE)
> freeparL[anchors1, 1] <- FALSE
> freeparL[anchors2, 2] <- FALSE
```

After running these lines of syntax, `freeparT` and `freeparL` are matrices with `TRUE` and `FALSE` entries indicating which intercepts and factor loadings, respectively, are allowed to differ as a function of the background variables.

In the anchors-only model, all studied indicators' intercepts and factor loadings are now allowed to differ as a function of the background variables. In order to specify the anchors-only model, the user should indicate which elements in `MxMatrix` objects `matB1`, `matB2`, `matC1`, and `matC2` can be freely estimated:

```
> matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=freeparT,
>                    values=0,
>                    name="matB1")
> matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=freeparT,
>                    values=0,
>                    name="matB2")
> matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                    free=freeparL,
>                    values=0,
>                    name="matC1")
> matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                    free=freeparL,
>                    values=0,
>                    name="matC2")
```

All other elements of the anchors-only model are similar to the all-but-one models specified in Step 4a. So after re-specifying `matB1`, `matB2`, `matC1`, and `matC2`, the anchors-only model can be fitted to the `DS14` data:

```
> modAnchors <- mxModel(model="AnchorsOnly",
>                       matT, matT0, matB1, matB2,
>                       matL, matL0, matC1, matC2,
>                       matE, matE0, matD1, matD2,
```

```
>                               matP, matP0, matH1, matH2,
>                               matA, matA0, matG1, matG2,
>                               matIa, matIb, matV1, matV2,
>                               matVar, matR, matCov, matM, matC,
>                               expF, fitF, mxdata1)
> fitAnchors <- mxRun(modAnchors)
```

The object `fitAnchors` contains the model fit and parameter estimates of the unconstrained model, which can be printed using the `summary()` function.

Step 5b: Specify and Fit Anchors-Plus-One Models

In each anchors-plus-one model, the studied indicator is additionally constrained to exhibit no DIF. That is, all intercepts and factor loadings are allowed to differ as a function of the background variables, except for the current studied indicator and the anchors. First, an empty list can be created for the output of all constrained models:

```
> fitApo <- list()
```

The anchors-plus-one model for each studied indicator can be specified and fit within this `for()` loop:

```
> for (i in testIn){
>     freeparTa <- freeparT
>     freeparLa <- freeparL
>     freeparTa[i] <- FALSE
>     freeparLa[i, ifelse(i < 8, yes=1, no=2)] <- FALSE
>     matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                       free=freeparTa,
>                       values=0,
>                       name="matB1")
>     matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                       free=freeparTa,
>                       values=0,
>                       name="matB2")
>     matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                       free=freeparLa,
>                       values=0,
>                       name="matC1")
>     matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                       free=freeparLa,
>                       values=0,
>                       name="matC2")
>     modApo <- mxModel(model=paste0("Anchors_plus_", i),
>                       matT, matT0, matB1, matB2,
>                       matL, matL0, matC1, matC2,
>                       matE, matE0, matD1, matD2,
>                       matP, matP0, matH1, matH2,
>                       matA, matA0, matG1, matG2,
>                       matIa, matIb, matV1, matV2,
>                       matVar, matR, matCov, matM, matC,
>                       expF, fitF, mxdata1)
>     fitApo[[i]] <- mxRun(modApo)
> }
```

So for each studied indicator, the matrices of freely estimated (non)uniform DIF (`freeparT` and `freeparL`) are copied in the first two lines of the `for()` loop, in order to additionally fix the DIF parameters of the current studied indicator `i` to zero in the following two

lines. After completing the `for()` loop, the object `fitApo` is a list containing the fit and parameter estimates of each studied indicators' anchors-plus-one model.

Step 5c: Conduct Likelihood-Ratio Tests

Partial scalar-and-metric invariance can now be assessed by performing a LRT for all studied indicators using the `mxCompare()` function:

```
> piTest <- mxCompare(fitAnchors, fitApo)
> piOut <- data.frame(Name  = paste0("Indicator", testIn),
>                      X2    = piTest$diffLL[2:11],
>                      df    = piTest$diffdf[2:11],
>                      p     = piTest$p[2:11],
>                      p.bon = p.adjust(p=piTest$p[2:11], method="bonferroni"),
>                      p.BH  = p.adjust(p=piTest$p[2:11], method="BH"))
```

The object `piOut` contains the LRT statistic, df, and p value for each studied indicator, presented in Table 5.2. To account for multiple testing, we also included Bonferroni-adjusted p values to control the familywise Type I error rate, as well as more-powerful Benjamini–Hochberg adjustments to maintain a false discovery rate no larger than the α level. Without accounting for multiple testing, the LRTs indicate that constraining the intercepts and factor loadings of Indicators 1, 2, 7, and 13 to be unaffected by `Male` or `Age` leads to a significantly worse model fit. Using $\alpha = .05$ as significance level, the H_0 of measurement invariance with respect to `Male` and `Age` for these indicators can thus be rejected. However, other conclusions can be made when accounting for multiple testing. The Bonferroni-adjusted p values indicate that only Indicator 13 significantly violates scalar-and-metric invariance, whereas the Benjamini–Hochberg-adjusted p values additionally indicate a significant violation of Indicator 7. Follow-up Wald tests of specific (non)uniform DIF can be conducted by consulting the Wald z statistics in the `summary()` output of the models with significant DIF. This may be warranted if more information about the nature of DIF is desired (e.g., to attempt revising indicators to remove such DIF). Our Open Science Framework project `https://osf.io/6cyxt/` also includes R code to inspect tracelines of the DIF indicators which may help with interpreting the DIF effects. In addition, R code for plots of moderated common-factor means, variances, and correlations can be found here along with image files for the plots themselves.

5.3.6   Final Model: Comparison with M*plus*

In this section of the tutorial, we fit the final partial-invariance model to the `DS14` data, using both `OpenMx` (Neale et al., 2016) and M*plus* (L. K. Muthén & Muthén, 2012). The purpose of this section is to show that the parameter estimates obtained by these two software packages are identical. In the final partial-invariance model, we assume scalar and metric invariance of all indicators except for Indicators 7 and 13. These indicators functioned differently with respect to the background variables, which is taken into account in the final partial-invariance model by freely estimating the effects of `Male` and `Age` on the intercept and factor loading of this indicator.

Table 5.2: Likelihood-ratio tests for assessing partial invariance.

Indicator	Name	$\Delta\chi^2(4)$	p	p_{bon}	$p_{ben-hoc}$
1	Si1	10.67	**.031**	.305	.089
2	Si3	10.30	**.036**	.357	.089
3	Si6	5.69	.224	1.000	.287
4	Si8	5.61	.230	1.000	.287
7	Si14	14.44	**.006**	.060	**.030**
10	Na5	8.50	.075	.748	.150
11	Na7	1.64	.802	1.000	.802
12	Na9	3.03	.554	1.000	.615
13	Na12	15.27	**.004**	**.042**	**.030**
14	Na13	5.91	.206	1.000	.287

Note. Indicators 5 and 6 are the anchors for social inhibition and Indicators 8 and 9 are the anchors for negative affectivity. The bold cells indicate significant (non)uniform DIF based on the original p value, the Bonferroni-adjusted p value denoted p_{bon}, and the Benjamini–Hochberg-adjusted p value denoted $p_{ben-hoc}$.

First, we can specify which indicators are scalar and metric invariant:

```
> finalIn <- c(1,2,3,4,5,6,8,9,10,11,12,14)
```

We can then use this `finalIn` object to indicate which effects of the background variables `Male` and `Age` on the intercepts and factor loadings should be freely estimated:

```
> freeparTb <- freeparT
> freeparLb <- freeparL
> for(i in finalIn){
>   freeparTb[1, i] <- FALSE
>   freeparLb[i, ifelse(i < 8, yes=1, no=2)] <- FALSE
> }
```

The matrices `freeparTb` and `freeparLb` indicate which DIF parameters should be estimated in the partial-invariance model. The matrices `matB1`, `matB2`, `matC1`, and `matC2` can now be re-specified:

```
> matB1 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                   free=freeparTb,
>                   values=0,
>                   name="matB1")
> matB2 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                   free=freeparTb,
>                   values=0,
>                   name="matB2")
> matC1 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                   free=freeparLb,
>                   values=0,
>                   name="matL1")
> matC2 <- mxMatrix(type="Full", nrow=nv, ncol=2,
>                   free=freeparLb,
>                   values=0,
>                   name="matC2")
```

and added to a new `MxModel` object `modOpenmxPartial`, which can then be fitted to data using the `mxRun()` function:

```
> modOpenmxPartial <- mxModel(model="PartialInvariance",
>                             matT, matT0, matB1, matB2,
>                             matL, matL0, matC1, matC2,
>                             matE, matE0, matD1, matD2,
>                             matP, matP0, matH1, matH2,
>                             matA, matA0, matG1, matG2,
>                             matIa, matIb, matV1, matV2,
>                             matVar, matR, matCov, matM, matC,
>                             expF, fitF, mxdata1)
> fitOpenmxPartial <- mxRun(modOpenmxPartial)
```

Again, the `summary()` function can be used to evaluate the fit of this model. The parameter estimates of this model are shown in the summary output and can be extracted with `summary(fitOpenmxPartial)$parameters`.

The same partial-invariance model can also be fitted to the data in M*plus* via the R package `MplusAutomation` (Hallquist & Wiley, 2018). Note that the R syntax below can only be executed if M*plus* is installed on the user's computer. The first step of fitting the partial-invariance model in M*plus* is to create a folder in which the empirical data and models can be stored:

```
> pathfix <- "~/FinalModel"
> dir.create(pathfix)
```

The `DS14` data can then be saved in this directory using the `prepareMplusData()` function:

```
> prepareMplusData(df=DS14, filename=paste0(pathfix, "/DS14dat.dat"))
```

Now that the `"~/FinalModel"` folder contains the empirical data, the partial-invariance model can be specified and added to the directory:

```
> modMplusPartial <- '
>     DATA:
>     FILE = "DS14dat.dat";
>
>     VARIABLE:
>     NAMES = Male Age x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14;
>     CONSTRAINT = Male Age;
>     MISSING = .;
>
>     ANALYSIS:
>     estimator is ML;
>
>     MODEL:
>     SI by x1-x7* (l1-l7);
>     NA by x8-x14* (l8-l14);
>     SI (SIvar);
>     NA (NAvar);
>     SI with NA (cov);
>     [SI] (SImean);
>     [NA] (NAmean);
>     [x1-x14] (t1-t14);
>     x1-x14 (e1-e14);
>     Male @1;
>     Age @1;
>     [Male @0];
>     [Age @0];
>
```

```
>       MODEL CONSTRAINT:
>       NEW (
>       h1_SI h1_NA g1_SI g1_NA
>       h2_SI h2_NA g2_SI g2_NA
>       p0 h1_cov h2_cov
>       t7_0 b7_1 b7_2 t13_0 b13_1 b13_2
>       l7_0 c7_1 c7_2 l13_0 c13_1 c13_2
>       e1_0 e2_0 e3_0 e4_0 e5_0 e6_0 e7_0 e8_0 e9_0 e10_0 e11_0 e12_0 e13_0 e14_0
>       d1_1 d2_1 d3_1 d4_1 d5_1 d6_1 d7_1 d8_1 d9_1 d10_1 d11_1 d12_1 d13_1 d14_1
>       d1_2 d2_2 d3_2 d4_2 d5_2 d6_2 d7_2 d8_2 d9_2 d10_2 d11_2 d12_2 d13_2 d14_2);
>
>       SIvar = 1 * EXP(h1_SI*Male + h2_SI*Age);
>       NAvar = 1 * EXP(h1_NA*Male + h2_NA*Age);
>       SImean = 0 + g1_SI*Male + g2_SI*Age;
>       NAmean = 0 + g1_NA*Male + g2_NA*Age;
>       cov = SQRT(EXP(h1_SI*Male + h2_SI*Age))*
>             SQRT(EXP(h1_NA*Male + h2_NA*Age))*
>             (EXP(2*(p0 + h1_cov*Male + h2_cov*Age))-1)/
>             (EXP(2*(p0 + h1_cov*Male + h2_cov*Age))+1);
>       e1 = e1_0 * EXP(d1_1*Male + d1_2*Age);
>       e2 = e2_0 * EXP(d2_1*Male + d2_2*Age);
>       e3 = e3_0 * EXP(d3_1*Male + d3_2*Age);
>       e4 = e4_0 * EXP(d4_1*Male + d4_2*Age);
>       e5 = e5_0 * EXP(d5_1*Male + d5_2*Age);
>       e6 = e6_0 * EXP(d6_1*Male + d6_2*Age);
>       e7 = e7_0 * EXP(d7_1*Male + d7_2*Age);
>       e8 = e8_0 * EXP(d8_1*Male + d8_2*Age);
>       e9 = e9_0 * EXP(d9_1*Male + d9_2*Age);
>       e10 = e10_0 * EXP(d10_1*Male + d10_2*Age);
>       e11 = e11_0 * EXP(d11_1*Male + d11_2*Age);
>       e12 = e12_0 * EXP(d12_1*Male + d12_2*Age);
>       e13 = e13_0 * EXP(d13_1*Male + d13_2*Age);
>       e14 = e14_0 * EXP(d14_1*Male + d14_2*Age);
>       t7 = t7_0 + b7_1*Male + b7_2*Age;
>       l7 = l7_0 + c7_1*Male + c7_2*Age;
>       t13 = t13_0 + b13_1*Male + b13_2*Age;
>       l13 = l13_0 + c13_1*Male + c13_2*Age;'
> cat(modMplusPartial, file = paste0(pathfix, "/modPartial.inp", sep = ""))
```

For a more detailed explanation of the M*plus* syntax, see Bauer's (2017) supplementary materials. For the sake of simplifying the comparison of results, we use `MplusAutomation` so that results can be imported into R, although R is not necessary to use M*plus* for MNLFA estimation. If one prefers to run the models in M*plus* directly instead of indirectly via `MplusAutomation`, the M*plus* script can be found on our Open Science Framework project `https://osf.io/6cyxt/`.

The `"/modPartial.inp"` file can be run and the corresponding results can be imported into R as follows:

```
> runModels(pathfix)
> fitMplusPartial <- readModels(pathfix)
```

The parameter estimates of this model stored in `fitMplusPartial$parameters` are identical to those obtained by `OpenMx`, as seen with `summary(fitOpenmxPartial)$parameters`. This provides a valuable cross-validation that the model is specified equivalently in both software packages and that both optimizers converge on the same full-information maximum likelihood estimates. Further research is needed to cross-validate across a wider variety of SEMs (e.g., categorical indicators; Bauer, 2017), but the current results imply that any

researcher can utilize MNLFA with the freely available `OpenMx` package, even if they do not have access to commercial software such as M*plus* or SAS.

5.4 Extensions

5.4.1 Nonlinear Effects Among Background Variables

One may want to add quadratic effects of a background variable, or in the case of MNLFA with multiple background variables, it may be desirable to account for the interactions between the background variables in their effects on factor-model parameters. Incorporating quadratic effects or interactions between background variables is straightforward. We simply need to calculate the quadratic or interaction variable in R and treat it as an additional moderator in `OpenMx`. For instance, if we are interested in the interaction between `Age` and `Male` in the real data example above, we can use:

```
> DS14$Int <- DS14$Age * DS14$Male
> mxdata_int <- mxData(observed=DS14, type="raw")
> matV2 <- mxMatrix(type="Full", nrow=1, ncol=1,
>                    free=FALSE,
>                    labels="data.Int",
>                    name="Int")
```

Now we can specify effects of this additional moderator along with the other effects on factor-model parameters—for example, the factor loadings:

```
> matB3 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                    free=freeparT,
>                    values=0,
>                    name="matB3")
> matT <- mxAlgebra(expression = matT0 + matB1*Male + matB2*Age + matB3*Int,
>                    name="matT")
```

That is, an additional parameter matrix, `matB3`, specifies the `Age` × `Male` interaction effects on factor loadings. We add these effects to the `expression` for the moderation of the factor loadings. For all other moderated parameters (intercepts, residual variances, factor means, and factor variances), the procedure is the same: an extra parameter matrix needs to be specified, and the interaction between the background variables needs to be added to the moderation function. Likewise, the quadratic effect of `Age` could be added by calculating a new variable that is the square of `Age`, then specifying additional parameter matrices for its effects.

5.4.2 Ordinal Data

If the indicators in the MNLFA model are ordinal, it may be advisable to explicitly treat them as ordinal (particularly with less than five categories) to prevent detection of spurious nonlinear effects. To do so, we must first specify that the data are ordinal. This can be done using the `mxFactor()` function:

```
> DS14[,3:16] <- mxFactor(DS14[,3:16], levels=0:4)
> mxdata_ord <- mxData(observed=DS14, type="raw")
```

```
> nc <- 5
```

Thus, we specified the indicators (i.e., the data in columns 3 to 16 of the DS14 matrix) to be ordinal with levels 0 to 4. Additionally, we assigned the data to object `mxdata_ord` and we specified an object (`nc`) to denote the number of categories per variable (five), which we use below to specify the model.

Next, to enable MNLFA of ordered categorical data, a discrete factor model is adopted for the data (for a formal description of such models, see B. Muthén, 1984; Takane & De Leeuw, 1987; Wirth & Edwards, 2007). In the discrete factor model, each observed ordinal indicator is assumed to have a corresponding normally distributed latent response variable (LRV) that is discretized by thresholds to yield the ordered categories. Therefore, to extend the code for the unconstrained model above, we specify a matrix containing the threshold parameters. If the ordinal indicators contain K categories, there are $K - 1$ threshold parameters per indicator. We therefore need a $(K - 1) \times P$ matrix in which rows represent thresholds and columns represent indicators:

```
> matThres <- mxMatrix(type="Full", nrow=nc-1, ncol=nv,
>                      free=TRUE,
>                      values=rep(seq(-2,2,length.out=nc-1),times=nv),
>                      byrow=FALSE,
>                      name="matThres")
```

Recall `nc` denotes a constant[6] number of categories per indicator (see above). The values for the thresholds can be negative, but should always be increasing for a given indicator. Therefore, as starting values, we specified increasing, equally spaced values between -2 and 2.

In the discrete factor model, the origin and scale of each LRV must be set, just as for latent common factors (Wu & Estabrook, 2016). In the code below, we fix the baseline intercepts to zero and baseline residual variances to one for all ordinal indicators, allowing us to freely estimate all[7] their thresholds.

```
> matT0 <- mxMatrix(type="Full", nrow=1, ncol=nv,
>                   free=FALSE,
>                   values=0,
>                   name="matT0")
> matE0 <- mxMatrix(type="Diag", nrow=nv, ncol=nv,
>                   free=FALSE,
>                   values=1,
>                   name="matE0")
```

Using the same `matThres` across all values of the background variables implies invariance of thresholds across the background variables, in which case the configuration of background-variable effects on all other model parameters need not change from those shown in Steps 4 and 5. We discuss this assumption of equal thresholds after fitting the model.

---

[6]In the case of an unequal number of categories across indicators, `nc` should be set to the maximum number of thresholds across all indicators. Thresholds that do not exist can be dropped from the model by specifying the `values=` argument to be `NA` (missing) and the `free=` argument as `FALSE` for nonexisting thresholds.

[7]Alternatively, we could fix two thresholds per indicator (e.g., to zero and one) to freely estimate the baseline intercepts and residual variances (see Mehta et al., 2004; Wu & Estabrook, 2016), which is consistent with the LISREL approach (Millsap & Tein, 2004).

Then we need to re-specify the expectation function to indicate the presence of thresholds in the model:

```
> expF <- mxExpectationNormal(covariance="matC",
>                             means="matM",
>                             thresholds="matThres",
>                             dimnames=manVars)
```

Finally, we can combine all objects into an `OpenMx` object and add the restriction that the thresholds are strictly increasing for a given indicator:

```
> modUn_thres <- mxModel(model="ThresholdInvariance",
>                        matT, matT0, matB1, matB2,
>                        matL, matL0, matC1, matC2,
>                        matE, matE0, matD1, matD2,
>                        matP, matP0, matH1, matH2,
>                        matA, matA0, matG1, matG2,
>                        matIa, matIb, matV1, matV2,
>                        matVar, matR, matCov, matM,
>                        matC, matThres, expF, fitF, mxdata_ord)
> modUn_thres_con <- mxConstrainMLThresholds(modUn_thres)
```

We can then fit the model

```
> fitUn_thres <- mxRun(modUn_thres_con)
```

Note that the threshold structure discussed above makes the model more complex as compared to an MNLFA model for continuous indicators. Therefore, the computation time for ordinal MNLFA models increases substantially, especially in the case of many indicators. In addition, optimization of the likelihood function is numerically more challenging which may cause the estimation to fail. For ordinal indicators, it is therefore advisable to use different starting values using the `mxTryHardOrdinal()` function. As compared to M*plus*, `OpenMx` is slower in the case of ordinal indicators. Fitting the final model above to the indicators of the SI factor only takes up to 70 minutes in `OpenMx` while it takes about a minute in M*plus* (on an Intel Core i5 with 8GB of RAM-memory). The advantage of `OpenMx` is however, that it is more flexible for models with ordinal indicators. For instance, residual variances can be moderated in `OpenMx` which is not possible in M*plus*.

The analysis above assumes rather than tests invariance of thresholds. We could test that assumption[8] by comparing its fit to an MNLFA in which thresholds (but not intercepts or residual variances) are functions of background variables. But when $K = 3$, these models would be statistically equivalent because (effects on) two thresholds are interchangeable with (effects on) the intercept and residual variance, so threshold invariance can only be assumed and not tested (Wu & Estabrook, 2016). In the special case of binary indicators, fixing effects on only one threshold cannot identify effects on both the intercept and residual variance. Thus, threshold invariance still cannot be independently tested, but neither can invariance of factor loadings and intercepts (Wu & Estabrook,

---

[8]Note that although it is popular to test invariance of ordinal indicators by leaving intercepts fixed to zero and testing threshold invariance in place of intercept invariance (i.e., after testing invariance of loadings; Millsap & Tein, 2004; Liu et al., 2017), Wu & Estabrook (2016) explained why this leads to invalid comparisons (i.e., the response scales are not linked unless ≥ 2 thresholds are equivalent across background variables).

2016). So for binary indicators, one could compare the fit of a configural model (with background-variable effects on factor loadings and intercepts, but not thresholds or residual variances) to a scalar-invariance model (with background-variable effects on residual variances, but not on thresholds, factor loadings, or intercepts). If H_0 can be rejected, partial invariance can still be established, but the data could not distinguish the nature of an indicator's DIF.

5.5   Discussion

This chapter illustrated how to perform MNLFA for measurement invariance assessment using the R package `OpenMx`. We considered a two-factor MNLFA model for the DS14 data (Denollet et al., 2013) and showed how to evaluate full and partial measurement invariance with respect to a dichotomous and continuous background variable simultaneously. This is one of the first papers showing how to test for measurement invariance with MNLFA in free and open-source SEM software, cross-validating its results with the M*plus* package. We therefore hope that with this tutorial we provide more researchers the opportunity to perform MNLFA for assessing measurement invariance or other purposes.

There are multiple advantages of MNLFA over other methods for testing measurement invariance. Unlike the MGCFA approach, MNLFA allows for the assessment of measurement invariance with respect to multiple background variables simultaneously. In addition, whereas MGCFA is only appropriate for categorical background variables, MNLFA also permits the assessment of measurement invariance with respect to continuous background variables. Whereas other single-group approaches (e.g., RFA and MIMIC) share these advantages, MNLFA does not require assuming common-factor or residual homoskedasticity with respect to the background variables. Although the use of product indicators to model moderation of factor loadings in RFA or MIMIC can capture some heteroskedasticity present in the data (Kolbe et al., 2021), it is not as flexible or interpretable as with MNLFA. We aimed to highlight all of these advantages in this tutorial by fitting an MNLFA model to empirical data that included all of these elements.

One of the steps included in this tutorial is selecting a set of anchor indicators. This step is useful for almost all methods for evaluating measurement invariance. A traditional strategy is to use all indicators other than the studied indicator as anchors, which leads to a contaminated set of anchors when one or more indicators in the anchor set violate measurement invariance. This can in turn lead to biased parameter estimates and inflated Type I error rates when assessing measurement invariance (W.-C. Wang, 2004). It is therefore advisable to use an anchor-selection strategy to select a smaller subset of anchor indicators (preferably 10–20% of the total number of indicators), such as the rank-based strategy proposed by Woods (2009). Several simulation studies have shown that this strategy frequently obtains uncontaminated sets of anchor indicators (Woods, 2009; M. Wang & Woods, 2017; Kolbe & Jorgensen, 2019; Kopf et al., 2015b). In this tutorial, we provided a step-by-step explanation of how to select anchor indicators using

the rank-based strategy. If the user wants to prevent making a decision about which indicators should serve as anchors, combining MNLFA with a regularization approach may be an appealing method to assess measurement invariance (see Bauer et al., 2020).

The key element of the MNLFA approach is that CFA parameters are predicted by the background variables. As such, a functional form has to be assumed between the parameters and background variables. The present tutorial illustrated how to model (log-)linear relationships between the background variables and model parameters in MNLFA, which might not always be an accurate representation of the data. In such situations, a researcher could consider higher-order polynomial functions (Bauer & Hussong, 2009) or semiparametric MNLFA approaches including the local SEM approach by Hildebrandt et al. (2016), the mixture approach by Molenaar (2020), or the score based approach by Merkle & Zeileis (2013). The advantage of these semiparametric MNLFA approaches is that an assumption about the functional form between the background variables and parameters is not required, making it a suitable approach for exploratory situations in which there is no theory about the functional form of the relationship.

One of the limitations of the current chapter is that we only demonstrated one method for assessing measurement invariance. Although this method seems to perform well across various conditions, many other methods are available as well. Therefore, we close by briefly mentioning the latest developments regarding methods for measurement invariance assessment and (open-source) software packages. From the semiparametric approaches discussed above, the local SEM approach can be applied using open-source R package `sirt` (Robitzsch, 2020a) and the score-based approach can be applied in open-source R package `lavaan` (Rosseel, 2012). The mixture approach is currently not implemented in an open-source package yet, but it can be applied using `OpenMx` in principle and it can readily be applied in M*plus* (L. K. Muthén & Muthén, 2012). Examples of related measurement-invariance tools available in open-source software packages include the automated multi-group tests from the R package `semTools` (Jorgensen et al., 2019) and the multi-group DIF tests for categorical data from R packages `difR` (Magis et al., 2010) and `mirt` (Chalmers, 2012). In addition, the cluster-based approach to identify anchor items to test for measurement invariance with respect to a continuous variable (Schulze & Pohl, 2021) can be applied within R using the MNLFA implementation from the present chapter.

Chapter 6

General Discussion

6.1   General Remarks

The measurement of latent constructs such as cognitive ability, attitudes, and beliefs, serves an important role in social and behavioral science research and applications. As such constructs cannot be measured directly, observed measures function as indicators of the latent construct. In order to meaningfully compare a latent construct across groups, each observed indicator must relate to the latent construct in the same way for all individuals or groups. This condition is also referred to as measurement invariance. Measurement invariance is a prerequisite for meaningful comparisons across individuals or groups on latent constructs. If measurement invariance with respect to a specific background variable holds, the measurement of the latent construct is invariant across that background variable. But if measurement invariance does not hold, differences in observed scores across individuals or groups may arise from differences on the background variable instead of differences on the latent construct. It is therefore important to assess measurement invariance before comparing individuals or groups on latent constructs.

A common class of methods for assessing measurement invariance within the structural equation modeling (SEM) framework is confirmatory factor analysis (CFA). Measurement invariance can be assessed in CFA models by means of a comparison of specific features of the model across different levels of the background variable. One of the traditional CFA methods to evaluate measurement invariance across a categorical background variable (e.g., group membership) is multiple-group CFA (MGCFA; Vandenberg & Lance, 2000). In MGCFA, a CFA model is estimated for each group separately and measurement invariance is assessed by comparing the fit of models with and without increasingly restrictive equality constraints on the measurement parameters across the groups. Full invariance can be examined with an omnibus test for a particular level of measurement invariance for all indicators simultaneously (Drasgow & Kanfer, 1985; Horn & McArdle, 1992; Finch & French, 2018; Marsh, 1994). When the omnibus null hypothesis of full invariance is rejected, partial invariance can be assessed with an omnibus test for each indicator separately. Establishing partial invariance requires comparing the fit of a model with and without equality constraints on the studied indicator's parameters while holding a subset of other indicators invariant across the groups. These latter indicators are also called anchor indicators and can be selected using an anchor-selection strategy (see Kopf et al., 2015a).

In addition to MGCFA, single-group methods have been proposed for the purpose of assessing measurement invariance, including restricted factor analysis (RFA; Oort, 1992), multiple indicator multiple cause (MIMIC; Jöreskog & Goldberger, 1975), and moderated nonlinear factor analysis (MNLFA; Bauer & Hussong, 2009). These single-group methods involve fitting a single CFA model to the data aggregated over all levels of the background variable and are therefore more suitable for smaller samples sizes, continuous background variables, more complex functional relationships between the background variables and the observed indicators, and testing for measurement invariance across multiple continuous

and categorical background variables simultaneously. Although research on single-group methods to assess measurement invariance is increasing (see Barendse et al., 2010, 2012; Bauer et al., 2020; Woods & Grimm, 2011), there are several unsolved issues and unstudied topics of research. For example, it is unknown how different approaches to single-group methods compare in certain situations. The performance of these methods thus remains subject of ongoing research.

The current dissertation focused on the performance of novel ways of assessing measurement invariance using single-group methods. In **Chapter 1**, the concept of measurement invariance was introduced. Then, the use of product indicators (PI) in RFA models was proposed and illustrated in **Chapter 2**. The performance of the PI method in RFA models was investigated more extensively with a simulation study presented in **Chapter 3** in which PI was compared to the more traditional latent moderated structural equations (LMS) method. In **Chapter 4**, the impact of common-factor and residual heteroskedasticity on the performance of RFA with LMS, RFA with PI, and the recently introduced MNLFA method was examined. Finally, **Chapter 5** included a demonstration of how measurement invariance can be assessed with MNLFA using open-source statistical software. In the next sections of the present chapter, the main findings of the abovementioned chapters are summarized, and their implications for future research are discussed.

6.2   Summary of Main Findings

The single-group method RFA is readily suited to assess scalar invariance, but requires an extended method to evaluate metric invariance. The extended method should enable modeling the latent interaction between the latent construct and the background variable. RFA is most commonly extended with LMS. Although LMS has shown to have high power to detect violations of metric invariance, severely inflated Type I error rates have also been observed when using this method in RFA (and statistically equivalent MIMIC) models (see Barendse et al., 2010, 2012; Woods & Grimm, 2011). Therefore, PI was proposed as an alternative to LMS in RFA models in **Chapter 2**. Using a single simulated dataset, this chapter showed how the PI method can be used in RFA models to assess metric invariance. We compared the conclusions with those reached using LMS in RFA models and found comparable results, which indicates that the PI method is a viable alternative to LMS. Because RFA with LMS can only be implemented in commercial software M*plus*, knowing that PI is a viable alternative to LMS provides more researchers the opportunity to assess metric invariance with RFA using any SEM software package.

The performance of PI in RFA models was investigated more extensively with two simulation studies presented in **Chapter 3**. The first simulation study focused on methods of empirically selecting anchor indicators for RFA models. Anchor indicators are required for linking the metric of the common factors across the background variables when assessing different levels of partial invariance and can be selected using any selection strategy. This simulation study compared two empirical anchor-selection strategies:

the rank-based strategy proposed by Woods (2009) and an iterative selection procedure proposed by Barendse et al. (2012). The results of the simulation revealed that the rank-based strategy had the lowest risk and degree of contamination in the subset of anchor indicators. This anchor-selection strategy outperformed the iterative selection procedure across each sample-size and DIF-magnitude condition.

In the second simulation study of this chapter, the PI and LMS methods in RFA models were extensively compared. Specifically, the Type I error rates and power of the LMS and PI methods to detect violations of scalar and metric invariance were evaluated. The performance of these methods was assessed in two scenarios: a best-case scenario and an empirical-selection scenario. In the best-case scenario, two known DIF-free indicators were used as anchor indicators and were not tested for DIF and in the empirical-selection scenario, the rank-based strategy was used to select anchor indicators. The PI method appeared to have similar power but lower Type I error rates compared to LMS in almost all conditions of both scenarios. These results indicate that using PI in RFA models can minimize the inflated Type I error rates obtained with LMS. Although it has been argued that the inflated Type I error rates observed with LMS might be caused by a contaminated set of anchor indicators (Woods, 2009), our results contradict this possible explanation. The severely inflated error rates were not only observed in the empirical scenario, but also in the best-case scenario with a DIF-free anchor set. This suggests that a contaminated anchor set may not fully account for the frequently observed inflated error rates when using LMS.

Another possible explanation for the inflated Type I error rates observed with LMS in RFA models is a violation of the assumption of common-factor homoskedasticity and residual homoskedasticity (Chun et al., 2016; Meredith & Teresi, 2006). In order to confirm or disconfirm this possible explanation, the impact of violations of these assumptions on the performance of RFA combined with LMS and PI was investigated in **Chapter 4**. In contrast to RFA, the recently proposed MNLFA (Bauer & Hussong, 2009; Bauer, 2017) method for assessing measurement invariance does not require assuming common-factor or residual homoskedasticity with respect to the background variable. Hence, a comparison between RFA and MNLFA under each of the different simulation conditions was included as well. The results of the simulation study presented in this chapter showed that the Type I error rates obtained by RFA/LMS substantially increased as a function of common-factor heteroskedasticity, whereas MNLFA and RFA with PI appeared to be robust against violations of common-factor homoskedasticity. These results were as expected, because MNLFA explicitly accounts for common-factor heteroskedasticity by allowing for an effect of the background variable on the common-factor variance and RFA with PI includes a covariance between the common factor and the interaction factor which indirectly captures information about the difference in common-factor variances across the background variable. In conditions with residual heteroskedasticity, somewhat different patterns of results were found. The Type I error rates were close to the nominal level of significance for MNLFA, slightly inflated for RFA with PI, and severely inflated

for RFA with LMS.

Given its flexibility and good performance shown in multiple simulation studies (see Bauer et al., 2020; Kolbe et al., 2021), MNLFA seems to be a promising method for assessing measurement invariance with respect to categorical and continuous background variables. Performing MNLFA for measurement invariance assessment may, however, not be straightforward for researchers without access to M*plus* or SAS. In **Chapter 5**, the accessibility of MNLFA was increased by providing a detailed guideline on performing this method in the open-source R (R Core Team, 2021) package `OpenMx` (Boker et al., 2011). The chapter included a demonstration of how MNLFA can be applied in R for evaluating full and partial measurement invariance with respect to a dichotomous and continuous background variable simultaneously, including a step-by-step explanation of how to select anchor indicators using the rank-based strategy (Woods, 2009). In addition to this demonstration, we compared the results with those obtained when using MNLFA in M*plus* and found identical parameter estimates. This provided a valuable cross-validation that the model is implemented similarly in the two software packages and that both optimizers converge on the same full-information maximum likelihood estimates.

6.3 Future Research Directions

The main findings of this dissertation have several implications for future studies. Although the aim was to examine the performance of single-group methods for the assessment of measurement invariance as extensively as possible, there are still several unanswered questions. Below, I elaborate on the future directions that are important in this line of research. I explain why these matters would be interesting to investigate and how it relates to the main findings of this dissertation as well as the findings of other researchers.

6.3.1 Other Simulation Conditions

One of the restrictions of this dissertation is the limited number of simulation conditions. The aim was to include the majority of relevant varying factors in the simulation designs, but there is a possibility that some other factors that may be relevant too have been excluded. For example, this dissertation only investigated the performance of unidimensional RFA and MNLFA models. **Chapter 5** did include a demonstration of a two-factor MNLFA with multidimensional data, but future simulation studies could investigate the behavior of multidimensional RFA and MNLFA models more extensively. This is especially interesting given the increasing complexity of these models when three or more common factors are present. A specific unanswered question is how to ensure the positive definiteness of the common-factor covariance matrix for MNLFA models with more than two common factors. Previous studies on MNLFA (Bauer, 2017; Bauer et al., 2020; Kolbe et al., 2021) applied an elementwise approach to estimate the moderated covariances between common factors. This approach imposes bounds of -1 and 1 on

the common-factor correlations, but with more than two common factors there are no restrictions ensuring a positive definite common-factor covariance matrix at all levels of the background variables. Future research is warranted to examine and compare different ways of ensuring positive definiteness of the common-factor covariance matrix of MNLFA models with more than two common factors.

In line with other studies (e.g., see Barendse et al., 2010, 2012), this dissertation has primarily focused on RFA and MNLFA for continuous indicators. Because tests and questionnaires in the field of social and behavioral sciences also commonly include binary and ordinal items, it would be interesting to study the performance of these single-group methods with binary and ordinal indicators. The MNLFA method and the RFA method combined with LMS can handle binary and ordinal indicators, but there are only few studies on the performance of these methods for detecting DIF in binary and ordinal indicators. Exceptions include studies by Woods & Grimm (2011) and Chun et al. (2016) showing that RFA/LMS models for binary and ordinal indicators obtain severely inflated Type I error rates when common-factor variances differ across the background variable, which is in line with our findings (see **Chapter 4**). Another exception includes a simulation study by Bauer et al. (2020) showing that a regularized MNLFA approach for binary indicators performs well in larger samples (e.g., $N = 2000$). Future research could focus on studying the performance of MNLFA models for DIF detection in ordinal indicators and increasing the computational efficiency of open-source software packages for estimating MNLFA models for ordinal indicators. An alternative single-group method that is potentially less computationally intensive is RFA with PI. A generalization of RFA/PI for binary and ordinal indicators is, however, less straightforward. Suppose the indicators of the common factor and the background variable are both ordinal. In such a situation, ideal indicators of the latent interaction factor are products of the latent continuous responses that are assumed to underlie the ordinal indicators. Lodder et al. (2019) investigated the performance of treating ordinal indicators as continuous to utilize PI, which brings up the question of how such product indicators can be interpreted. The use of product indicators for the specific purpose of measurement invariance assessment with ordinal data is yet unexplored. Hence, much more research is needed in this area.

Another largely undeveloped area of research is the performance of single-group methods in conditions in which the assumption of multivariate normality of the observed variables is violated. This dissertation already studied the impact of a violation of homoskedasticity on assessing measurement invariance in single-group models like RFA, but what is the impact of a violation of the assumption of multivariate normality? The problem of nonnormality is not unique to measurement invariance (see Curran et al., 1996; Finney & DiStefano, 2006), but applies to many SEM applications because it is an assumption of the maximum likelihood estimator. Although the impact of nonnormality on measurement parameters in a factor model has been studied outside the context of measurement invariance, it is unknown how such a violation affects the assessment of measurement invariance with single-group methods. In a more general context, previous

studies have shown that fitting factor models to nonnormal data can result in underestimated standard errors, which leads to null hypotheses of individual parameters being rejected too often. Therefore, inflated Type I error rates may be observed when assessing measurement invariance using RFA or MNLFA in nonnormal conditions. The impact of nonnormality on the performance of the MGCFA method (Finch et al., 2018) and likelihood ratio method in item response theory (IRT; Woods, 2008) has already been studied. The findings of these studies were contradictory, as Finch et al. (2018) observed well-controlled Type I error rates while Woods (2008) observed inflated Type I error rates. Future research could focus on the effect of nonnormality on the performance of RFA and MNLFA, including a wider range of skewness and kurtosis conditions.

6.3.2 Other Methods for Assessing Measurement Invariance

This dissertation includes studies on three different methods to assess measurement invariance, but measurement invariance can be evaluated in many other ways. To begin with, this dissertation only considered one out of various PI approaches for estimating the latent interaction between the common factor and background variable in RFA models. We had a clear rationale for using the double-mean-centering strategy (Lin et al., 2010) because it eliminates the need to estimate a mean structure and does not require using a cumbersome multistage estimation procedure. An additional advantage of this approach is its robustness to nonnormality, but other PI approaches such as the orthogonalizing approach (see Little et al., 2006) may perform just as well or even better in the context of measurement invariance. Several other aspects of the use of PI are yet unclear, for example, which indicators should be used to build product indicators. There are multiple possibilities regarding the formation of product indicators, among which is using all indicators of the common factor and the background variable, as employed in this dissertation. But what would be a suitable strategy if, for example, both the common factor and the background variable are measured with 10 indicators? Building all possible product indicators may then no longer be a suitable strategy, as the latent interaction factor would end up with 100 indicators. More research is needed to determine the optimal use of PI in RFA models for the purpose of assessing measurement invariance.

With regards to MNLFA, a functional form has to be assumed between the model parameters and background variables. This dissertation only included an illustration of modeling linear and log-linear relationships between the background variable and parameters in MNLFA, but note that other functional forms may sometimes represent the data more accurately. A researcher could, for example, also consider higher-order polynomial functions or functions including interactions across multiple background variables (see Bauer & Hussong, 2009; Bauer et al., 2020). An alternative is to apply a semiparametric MNLFA like the mixture approach by Molenaar (2020). The advantage of this semiparametric MNLFA approach is that an assumption about the functional form between the background variable and model parameters is not required. Such an approach is thus

suitable for situations in which a researcher has little to no theory yet about the functional form of the relationship. The mixture approach is currently not implemented in open-source statistical software yet, but could be applied in `OpenMx` (Boker et al., 2011) or M*plus* (L. K. Muthén & Muthén, 2012).

In addition to the single-group methods investigated in this dissertation, there may be other families of methods for assessing measurement invariance that are more powerful or less sensitive to Type I error rates. One of these methods is SEM trees (Brandmaier et al., 2013) which allows for the detection of differences in parameter estimates across continuous or categorical background variables by recursively partitioning the data into subsets with significantly different SEM-parameter estimates. Simulation studies showed that SEM trees are capable of correctly partitioning the data into subsets with different parameter estimates (Usami et al., 2017, 2019) and detecting violations of scalar invariance in IRT models (Tutz & Berger, 2016; Strobl et al., 2015) in large samples. The performance of SEM trees in smaller samples and its power to detect violations of metric invariance has yet to be investigated.

Other methods for the assessment of measurement invariance worth mentioning are local SEM (Hildebrandt et al., 2016), heteroskedastic latent trait models (Molenaar et al., 2012; Molenaar, 2015; Molenaar et al., 2011; Molenaar, Dolan, & Verhelst, 2010), and stochastic process-based testing (Merkle & Zeileis, 2013; Merkle et al., 2014). Similar to RFA and MNLFA, these methods do not require splitting the data into groups before the models to the data and may therefore be more suitable than the MGCFA method for continuous background variables. Additional simulation studies could explore the power and Type I error rates of these alternative methods in comparison to RFA and MNLFA. From the methods discussed above, the local SEM method is available in the open-source R package `sirt` (Robitzsch, 2020a) and the score-based method is available in the open-source R package `lavaan` (Rosseel, 2012).

6.4 Concluding Remarks

This dissertation focused on novel approaches to assess measurement invariance using SEM. In specific, the research of this dissertation was aimed at investigating methods that are suitable for situations with continuous background variables and relatively small sample sizes, both being factors that are commonly present in the field of social and behavioral sciences (for examples, see Cheng & Watkins, 2000; Goodrich & Ercikan, 2019; Sudarshan et al., 2016; de Frias & Dixon, 2005). The findings of this dissertation suggest that MNLFA and RFA combined with PI are suitable methods for the assessment of measurement invariance in such situations, while RFA combined with LMS comes with convergence issues as well as a higher risk of falsely detecting violations of measurement invariance. We therefore recommend the use of MNLFA and RFA with PI over RFA with LMS in situations that are comparable with our simulation conditions. By showing how the well-performing methods can be applied with open-source statistical software, the

aim was to make these methods as accessible as possible to a wide range of researchers. This dissertation contributes to information about how best to evaluate measurement invariance, which may eventually lead to more valid research including latent constructs and fairer comparisons or decisions based on the measurement of latent constructs in practice. Because some issues still haunt the assessment of measurement invariance, we hope to inspire other researchers to continue this line of research.

Appendices

Appendix I: The Covariance between T and $T \times V$

Consider the one-factor model for the common factor T given by

$$x_{ip} = \tau_p + \lambda_p t_i + \varepsilon_{ip}, \qquad (A1)$$

where x_{ip} is the observed indicator score of person $i = 1, \ldots, N$ on indicator $p = 1, \ldots, P$, τ_p is an intercept, λ_p is a factor loading, t_i is a common-factor score and ε_{ip} is a residual. In addition, consider a background variable V to be a grouping variable dummy-coded $v_i = 0, 1$, representing membership in a reference or focal group, respectively. In this proof V is a categorical variable, but the proof generalizes to a continuous V.

Below we demonstrate that group differences in $\mathrm{Var}(T)$ can be captured by the interaction between T and V.

Let σ^2 denote the variance of the common factor T. First, we specify T as a scaled version of T', which has unit-variance:

$$T = \sigma T', \qquad (A2)$$

where

$$\mathrm{Var}(T') = 1. \qquad (A3)$$

A traditional two-group factor model with unequal variances in T between the groups can be written as

$$x_{ip} = \tau_p + \lambda_p \sigma t'_i + \varepsilon_{ip}, \qquad (A4)$$

where

$$\sigma = \sigma_0 + \sigma_1 V. \qquad (A5)$$

In this model, $\mathrm{Var}(T|v_i = 0) = \sigma_0^2$ and $\mathrm{Var}(T|v_i = 1) = (\sigma_0 + \sigma_1)^2$ which is equivalent to a two-group one-factor model with equal factor loadings, residual variances, and intercepts, but with unequal variance of T across groups.

Substituting Equation A5 in Equation A4 and slightly rewriting, we obtain

$$x_{ip} = \tau_p + \lambda_p (\sigma_0 t'_i + \sigma_1 v_i t'_i) + \varepsilon_{ip}, \qquad (A6)$$

which is the one-factor measurement model from Equation A1, but with the common factor T from Equation A1 regressed on VT' in the structural model.

The proof that a covariance between T and VT captures the information in σ_1 is that σ_1 is the effect of VT' on T. In simple regression, a slope is a simple function of the

analogous covariance and variance of the predictor:

$$\beta_{Y,X} = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}. \tag{A7}$$

Then it would follow from Equation A7 and Equation A8 treating T as Y and VT' as X that

$$\sigma_1 = \frac{\mathrm{Cov}(T,VT')}{\mathrm{Var}(VT')}. \tag{A8}$$

However, because Equation A7 is not analogous to a simple regression model but a multiple regression, expressing σ_1 as a function of $\mathrm{Cov}(T,VT')$ would be more complicated:

$$
\begin{aligned}
\sigma_1 &= \frac{\mathrm{Cov}(T,VT')\mathrm{Var}(T') - \mathrm{Cov}(T,T')\mathrm{Cov}(T',VT')}{\mathrm{Var}(VT')\mathrm{Var}(T') - \mathrm{Cov}(VT',T'))^2} \\
&= \frac{\mathrm{Cov}(T,VT') - \mathrm{Cov}(T,T')\mathrm{Cov}(T',VT')}{\mathrm{Var}(VT') - \mathrm{Cov}(VT',T'))^2}.
\end{aligned}
\tag{A9}
$$

Replacing T' by $\sigma^{-1}T$, the expression of σ_1 in Equation A9—which is the difference in common-factor variances across groups—is a complex function of three model parameters: the variances of the common factor and interaction terms and their covariance.

$$
\begin{aligned}
\sigma_1 &= \frac{\mathrm{Cov}(T,V\sigma^{-1}T) - \mathrm{Cov}(T,\sigma^{-1}T)\mathrm{Cov}(\sigma^{-1}T,V\sigma^{-1}T)}{\mathrm{Var}(V\sigma^{-1}T) - \mathrm{Cov}(V\sigma^{-1}T,\sigma^{-1}T))^2} \\
&= \frac{\sigma^{-1}\mathrm{Cov}(T,VT) - \sigma^{-3}\mathrm{Cov}(T,VT)}{\sigma^{-1}\mathrm{Var}(VT) - \sigma^{-4}\mathrm{Cov}(VT,T))^2}.
\end{aligned}
\tag{A10}
$$

Because a regression slope (or a correlation) between two variables is simply a ratio of their covariance to the variance of the predictor (or to the product of their standard deviations), it follows that by estimating the parameters $\mathrm{Cov}(T,VT)$, $\mathrm{Var}(VT)$, and $\mathrm{Var}(T) = \sigma^2$, RFA models with product indicators indirectly capture the same information about common-factor heteroskedasticity that MNLFA can capture by directly estimating the slope σ_1.

Appendix II: Supplementary Tables

Table A1: The population parameter values for each of the indicators across all conditions

	Continuous V			Categorical V		
Indicator	b	c	d	b	c	d
1	0	0	-0.25/0/0.25	0	0	$\ln(\frac{0.15}{0.3})/0/\ln(\frac{0.6}{0.3})$
2	0.25	0	-0.25/0/0.25	0.50	0	$\ln(\frac{0.15}{0.3})/0/\ln(\frac{0.6}{0.3})$
3	0.25	0	0	0.50	0	0
4	0	0.10	-0.25/0/0.25	0	0.25	$\ln(\frac{0.15}{0.3})/0/\ln(\frac{0.6}{0.3})$
5	0	0.10	0	0	0.25	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
10	0	0	0	0	0	0

Note. b is the effect of V on the indicator's intercept, c is the effect of V on the indicator's factor loading, and d is the effect of V on the indicator's residual variance.

Table A2: The nonconvergence of RFA/LMS across
all condition with a continuous V

h	d	N	Percentage of nonconvergence
-0.25	-0.25	100	2.50
		200	0.90
		500	0.00
		1000	0.00
	0	100	2.50
		200	0.70
		500	0.00
		1000	0.00
	0.25	100	4.80
		200	0.70
		500	0.00
		1000	0.00
0	-0.25	100	2.00
		200	0.70
		500	0.00
		1000	0.00
	0	100	2.90
		200	0.60
		500	0.00
		1000	0.00
	0.25	100	1.80
		200	0.60
		500	0.00
		1000	0.00
0.25	-0.25	100	3.40
		200	1.80
		500	0.10
		1000	0.00
	0	100	3.10
		200	1.30
		500	0.20
		1000	0.00
	0.25	100	3.00
		200	1.40
		500	0.00
		1000	0.00

Note. h = the effect of V on the common-factor variance,
d = the effect of V on the indicator's residual variance, N
= total sample size.

Table A3: The nonconvergence of RFA/LMS across all
condition with a categorical V

h	d	N	Percentage of nonconvergence
ln(0.5)	ln(0.15/0.3)	100	8.70
		200	2.20
		500	0.00
		1000	0.00
	0	100	12.10
		200	4.50
		500	0.70
		1000	0.00
	ln(0.6/0.3)	100	16.40
		200	10.00
		500	4.30
		1000	0.50
0	ln(0.15/0.3)	100	12.40
		200	5.00
		500	0.50
		1000	0.00
	0	100	12.70
		200	7.00
		500	1.10
		1000	0.00
	ln(0.6/0.3)	100	20.50
		200	15.70
		500	6.30
		1000	1.00
ln(1.5)	ln(0.15/0.3)	100	25.00
		200	19.60
		500	17.90
		1000	26.50
	0	100	24.30
		200	19.30
		500	25.10
		1000	38.90
	ln(0.6/0.3)	100	27.30
		200	26.50
		500	33.40
		1000	56.60
ln(2)	ln(0.15/0.3)	100	29.60
		200	24.50
		500	20.80
		1000	33.00
	0	100	27.40
		200	22.00
		500	20.70
		1000	35.40
	ln(0.6/0.3)	100	26.00
		200	22.70
		500	19.20
		1000	35.20

Note. h = the effect of V on the common-factor variance,
d = the effect of V on the indicator's residual variance, N
= total sample size.

Table A4: The power of each method across all condition with a continuous V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
-0.25	-0.25	100	.710	.627	.693	.660	.687	.655
		200	.925	.859	.920	.899	.921	.908
		500	1.000	1.000	1.000	1.000	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.707	.635	.694	.651	.694	.660
		200	.916	.839	.904	.886	.900	.888
		500	1.000	.998	1.000	1.000	1.000	.999
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0.25	100	.703	.582	.686	.637	.682	.646
		200	.924	.859	.910	.895	.907	.911
		500	1.000	.998	1.000	.999	1.000	.999
		1000	1.000	1.000	1.000	1.000	1.000	1.000
0	-0.25	100	.705	.704	.700	.644	.678	.647
		200	.913	.906	.899	.888	.891	.887
		500	1.000	1.000	1.000	1.000	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.699	.654	.685	.623	.679	.635
		200	.918	.914	.905	.885	.894	.904
		500	1.000	1.000	1.000	.999	1.000	.999
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0.25	100	.697	.697	.676	.650	.658	.655
		200	.914	.900	.899	.883	.889	.898
		500	.999	1.000	.999	.999	.999	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000
0.25	-0.25	100	.703	.736	.682	.626	.667	.623
		200	.921	.960	.909	.888	.893	.903
		500	1.000	1.000	1.000	.999	1.000	.999
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.706	.781	.672	.663	.652	.643
		200	.915	.947	.898	.880	.891	.890
		500	1.000	1.000	.999	1.000	.999	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0.25	100	.727	.758	.699	.658	.682	.631
		200	.931	.959	.918	.888	.921	.902
		500	1.000	1.000	1.000	1.000	1.000	1.000
		1000	1.000	1.000	1.000	1.000	1.000	1.000

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. Power was calculated for Indicator 2 (violating scalar invariance, that is, $b_2 \neq 0$) and Indicator 4 (violating metric invariance, that is, $c_4 \neq 0$).

Table A5: The Type I error rates of each method across all condition with a continuous V

h	d	N	RFA/LMS	RFA/PI	MNLFA
-0.25	-0.25	100	**.125**	**.093**	**.118**
		200	**.111**	.062	**.075**
		500	**.142**	**.067**	.052
		1000	**.177**	**.064**	.050
	0	100	**.114**	**.086**	**.130**
		200	**.116**	**.078**	**.087**
		500	**.101**	.056	.058
		1000	**.138**	.046	.054
	0.25	100	**.130**	**.103**	**.157**
		200	**.109**	**.065**	**.080**
		500	**.079**	.057	.062
		1000	**.098**	.062	.047
0	-0.25	100	**.114**	**.095**	**.108**
		200	**.089**	**.069**	**.081**
		500	**.081**	**.073**	**.067**
		1000	.059	.052	.046
	0	100	**.116**	**.098**	**.109**
		200	**.078**	**.070**	**.077**
		500	.061	.056	.054
		1000	.060	.054	.057
	0.25	100	**.120**	**.114**	**.136**
		200	**.105**	**.105**	**.095**
		500	**.071**	**.067**	.054
		1000	**.066**	**.067**	.060
0.25	-0.25	100	**.134**	**.109**	**.108**
		200	**.109**	**.078**	**.077**
		500	**.093**	.057	.060
		1000	**.090**	.052	.053
	0	100	**.118**	**.086**	**.097**
		200	**.129**	**.084**	**.085**
		500	**.110**	.058	.061
		1000	**.148**	.054	.055
	0.25	100	**.141**	**.090**	**.088**
		200	**.126**	**.076**	**.070**
		500	**.130**	.057	.061
		1000	**.184**	.061	.056

Note. $h =$ the effect of V on the common-factor variance, $d =$ the effect of V on the indicator's residual variance, $N =$ total sample size. Boldface cells indicate a significant inflation. The Type I error rates were calculated for Indicator 1.

Table A6: The relative bias of each method across all condition with a continuous V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
-0.25	-0.25	100	5.518	26.259	0.764	26.048	7.902	17.142
		200	3.019	26.211	-0.335	24.851	1.586	3.094
		500	3.323	26.872	1.091	25.145	1.087	-0.453
		1000	2.220	27.059	0.425	24.596	0.448	-0.321
	0	100	3.987	23.352	-0.498	25.352	10.293	22.102
		200	2.319	24.253	-1.085	24.766	0.025	0.833
		500	2.260	28.041	-0.173	26.665	-0.173	0.811
		1000	1.159	28.471	-0.808	26.420	-0.746	0.843
	0.25	100	3.185	21.929	-0.788	24.988	9.530	24.442
		200	3.817	25.765	-0.062	26.118	1.414	3.173
		500	1.710	26.138	-0.845	25.055	-0.776	-0.343
		1000	1.913	26.268	-0.040	24.469	-0.021	-0.431
0	-0.25	100	3.059	40.182	-0.732	26.023	2.227	2.868
		200	1.430	40.368	-1.031	25.777	-0.784	-1.161
		500	1.094	43.623	-0.151	26.527	-0.124	0.670
		1000	0.418	43.651	-0.786	26.480	-0.749	0.796
	0	100	2.530	37.695	-1.034	23.821	3.821	8.835
		200	2.433	42.449	-0.040	26.134	-0.182	-0.076
		500	1.147	41.708	-0.695	24.994	-0.659	-0.370
		1000	1.274	41.638	-0.022	24.962	0.009	-0.387
	0.25	100	3.320	40.278	-0.092	25.718	4.735	9.396
		200	2.105	38.274	-0.112	23.662	-0.191	-2.468
		500	2.208	41.259	0.176	24.892	0.224	-0.469
		1000	1.625	41.357	0.132	25.148	0.142	-0.414
0.25	-0.25	100	5.064	55.608	-1.028	24.135	1.105	2.216
		200	5.581	59.562	-0.047	25.954	-0.131	-0.207
		500	2.737	59.253	-0.854	25.072	-0.736	-0.450
		1000	2.802	59.296	0.005	24.980	0.053	-0.298
	0	100	6.407	57.640	0.053	24.838	2.297	2.418
		200	5.479	56.229	0.015	24.104	-0.053	-2.364
		500	4.377	58.755	0.131	24.970	0.099	-0.615
		1000	3.392	59.171	0.125	25.138	0.151	-0.380
	0.25	100	7.873	57.597	1.084	25.485	2.186	-0.097
		200	5.374	57.639	-0.186	24.564	-0.213	-1.433
		500	4.770	59.071	1.151	25.591	1.187	-0.402
		1000	3.434	59.183	0.320	24.999	0.386	-0.308

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The relative bias was calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A7: The RMSE of each method across all condition with a continuous V

			RFA/LMS		RFA/PI		MNLFA	
h	d	N	Scalar	Metric	Scalar	Metric	Scalar	Metric
-0.25	-0.25	100	0.106	0.066	0.092	0.063	0.139	0.114
		200	0.075	0.047	0.067	0.043	0.089	0.055
		500	0.047	0.036	0.044	0.034	0.044	0.017
		1000	0.031	0.032	0.031	0.030	0.031	0.012
	0	100	0.109	0.063	0.099	0.061	0.154	0.127
		200	0.076	0.047	0.068	0.044	0.081	0.049
		500	0.049	0.038	0.043	0.036	0.044	0.018
		1000	0.032	0.033	0.031	0.031	0.031	0.012
	0.25	100	0.104	0.065	0.097	0.061	0.157	0.138
		200	0.081	0.048	0.070	0.045	0.086	0.053
		500	0.048	0.036	0.042	0.034	0.042	0.018
		1000	0.031	0.031	0.030	0.029	0.030	0.012
0	-0.25	100	0.111	0.071	0.099	0.063	0.114	0.068
		200	0.073	0.057	0.069	0.046	0.073	0.033
		500	0.045	0.050	0.044	0.035	0.044	0.018
		1000	0.031	0.047	0.031	0.031	0.031	0.012
	0	100	0.103	0.071	0.096	0.060	0.124	0.093
		200	0.077	0.058	0.070	0.045	0.071	0.029
		500	0.048	0.049	0.042	0.034	0.042	0.018
		1000	0.030	0.045	0.030	0.030	0.030	0.012
	0.25	100	0.108	0.069	0.101	0.059	0.131	0.090
		200	0.074	0.055	0.070	0.044	0.070	0.028
		500	0.049	0.048	0.044	0.033	0.044	0.017
		1000	0.032	0.045	0.031	0.030	0.031	0.012
0.25	-0.25	100	0.110	0.082	0.096	0.061	0.109	0.068
		200	0.086	0.072	0.070	0.045	0.071	0.029
		500	0.049	0.064	0.042	0.034	0.042	0.018
		1000	0.031	0.062	0.030	0.030	0.030	0.012
	0	100	0.114	0.081	0.100	0.058	0.120	0.067
		200	0.086	0.069	0.069	0.044	0.070	0.028
		500	0.055	0.064	0.043	0.033	0.043	0.017
		1000	0.035	0.062	0.031	0.030	0.031	0.012
	0.25	100	0.109	0.083	0.093	0.062	0.099	0.052
		200	0.081	0.070	0.067	0.044	0.067	0.028
		500	0.050	0.064	0.044	0.034	0.044	0.018
		1000	0.032	0.062	0.031	0.030	0.031	0.013

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The RMSE was calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A8: The coverage rates of each method across all condition with a continuous V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
-0.25	-0.25	100	.929	.855	.946	.875	.935	.865
		200	.929	.840	.951	.886	.941	.922
		500	.927	.779	.940	.802	.940	.936
		1000	.934	.595	.946	.652	.945	.934
	0	100	.920	.875	.933	.879	.907	.873
		200	.932	.867	.942	.871	.933	.923
		500	.931	.752	.949	.756	.943	.926
		1000	.936	.570	.951	.616	.947	.938
	0.25	100	.922	.869	.947	.889	.908	.871
		200	.917	.866	.938	.872	.936	.916
		500	.944	.783	.952	.799	.953	.917
		1000	.936	.629	.952	.646	.952	.946
0	-0.25	100	.919	.835	.940	.877	.925	.899
		200	.925	.759	.942	.866	.940	.923
		500	.929	.520	.947	.766	.945	.937
		1000	.941	.241	.943	.624	.939	.944
	0	100	.925	.833	.934	.895	.921	.894
		200	.928	.763	.940	.875	.939	.919
		500	.944	.566	.954	.793	.955	.913
		1000	.947	.299	.954	.635	.956	.942
	0.25	100	.914	.842	.924	.890	.903	.903
		200	.932	.782	.943	.881	.938	.937
		500	.935	.554	.947	.812	.947	.940
		1000	.916	.314	.936	.643	.935	.947
0.25	-0.25	100	.918	.760	.939	.892	.930	.902
		200	.915	.621	.936	.873	.935	.929
		500	.946	.291	.952	.792	.951	.931
		1000	.934	.066	.959	.632	.949	.946
	0	100	.902	.756	.929	.888	.920	.913
		200	.919	.628	.942	.872	.936	.933
		500	.923	.295	.940	.797	.939	.938
		1000	.911	.075	.941	.639	.938	.948
	0.25	100	.925	.735	.949	.870	.946	.900
		200	.918	.623	.948	.880	.945	.936
		500	.912	.293	.935	.793	.935	.938
		1000	.921	.082	.943	.632	.939	.930

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The coverage rates were calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A9: The power of each method across all condition with a categorical V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
ln(0.5)	ln(0.15/0.3)	100	.966	.128	.959	.295	.954	.229
		200	.999	.185	.999	.455	1.000	.453
		500	1.000	.360	1.000	.847	1.000	.906
		1000	1.000	.657	1.000	.990	1.000	.998
	0	100	.936	.123	.916	.295	.909	.161
		200	1.000	.202	.999	.462	.999	.379
		500	1.000	.445	1.000	.888	1.000	.844
		1000	1.000	.748	1.000	.996	1.000	.995
	ln(0.6/0.3)	100	.865	.150	.837	.260	.826	.140
		200	.991	.229	.984	.485	.985	.312
		500	1.000	.484	1.000	.859	1.000	.721
		1000	1.000	.789	1.000	.988	1.000	.956
0	ln(0.15/0.3)	100	.963	.401	.958	.379	.956	.328
		200	1.000	.648	1.000	.599	1.000	.621
		500	1.000	.980	1.000	.953	1.000	.973
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.926	.447	.917	.400	.913	.295
		200	.999	.630	.999	.582	.999	.513
		500	1.000	.977	1.000	.955	1.000	.943
		1000	1.000	1.000	1.000	.999	1.000	.999
	ln(0.6/0.3)	100	.845	.397	.835	.355	.836	.240
		200	.987	.619	.988	.602	.981	.458
		500	1.000	.951	1.000	.931	1.000	.852
		1000	1.000	.999	1.000	.998	1.000	.997
ln(1.5)	ln(0.15/0.3)	100	.972	.615	.964	.451	.963	.384
		200	1.000	.884	1.000	.649	1.000	.658
		500	1.000	.998	1.000	.973	1.000	.981
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.950	.638	.931	.447	.927	.350
		200	1.000	.886	1.000	.697	1.000	.613
		500	1.000	.997	1.000	.979	1.000	.976
		1000	1.000	1.000	1.000	1.000	1.000	.999
	ln(0.6/0.3)	100	.861	.563	.840	.398	.839	.289
		200	.988	.872	.986	.695	.985	.560
		500	1.000	.994	1.000	.970	1.000	.920
		1000	1.000	1.000	1.000	1.000	1.000	1.000
ln(2)	ln(0.15/0.3)	100	.976	.753	.960	.460	.961	.408
		200	.999	.956	1.000	.719	1.000	.724
		500	1.000	1.000	1.000	.985	1.000	.994
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	0	100	.939	.733	.926	.501	.926	.406
		200	1.000	.945	.999	.708	.999	.653
		500	1.000	1.000	1.000	.981	1.000	.979
		1000	1.000	1.000	1.000	1.000	1.000	1.000
	ln(0.6/0.3)	100	.853	.691	.839	.448	.838	.325
		200	.991	.942	.990	.737	.985	.602
		500	1.000	1.000	1.000	.977	1.000	.943
		1000	1.000	1.000	1.000	1.000	1.000	1.000

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. Power was calculated for Indicator 2 (violating scalar invariance, that is, $b_2 \neq 0$) and Indicator 4 (violating metric invariance, that is, $c_4 \neq 0$).

Table A10: The Type I error rates of each method across all condition with a categorical V

h	d	N	RFA/LMS	RFA/PI	MNLFA
ln(0.5)	ln(0.15/0.3)	100	**.181**	**.064**	.052
		200	**.263**	**.065**	.056
		500	**.582**	**.068**	.052
		1000	**.865**	**.112**	**.073**
	0	100	**.127**	**.091**	**.076**
		200	**.155**	**.074**	**.069**
		500	**.294**	.061	.056
		1000	**.523**	.047	.047
	ln(0.6/0.3)	100	**.100**	**.084**	**.070**
		200	**.109**	**.069**	.063
		500	**.126**	**.079**	.060
		1000	**.198**	**.087**	.055
0	ln(0.15/0.3)	100	**.071**	**.069**	**.064**
		200	**.071**	**.076**	.062
		500	**.066**	**.069**	.048
		1000	**.069**	**.069**	.051
	0	100	**.077**	**.085**	**.081**
		200	.053	.054	.054
		500	.054	.045	.042
		1000	.057	.062	.058
	ln(0.6/0.3)	100	**.072**	**.076**	.055
		200	**.076**	**.076**	.063
		500	.054	.057	.050
		1000	**.077**	**.085**	.054
ln(1.5)	ln(0.15/0.3)	100	.057	.062	.059
		200	**.073**	**.064**	.060
		500	**.072**	.055	.053
		1000	**.109**	**.094**	**.068**
	0	100	**.083**	**.088**	**.083**
		200	**.092**	**.069**	.061
		500	**.121**	.053	.054
		1000	**.152**	.046	.044
	ln(0.6/0.3)	100	**.094**	**.081**	**.078**
		200	**.113**	**.076**	**.066**
		500	**.147**	**.066**	**.065**
		1000	**.276**	**.084**	.047
ln(2)	ln(0.15/0.3)	100	**.111**	**.073**	**.066**
		200	**.121**	**.072**	**.067**
		500	**.181**	.060	.053
		1000	**.309**	.061	.048
	0	100	**.123**	**.085**	**.080**
		200	**.108**	.060	.054
		500	**.241**	.049	.045
		1000	**.481**	.062	**.064**
	ln(0.6/0.3)	100	**.118**	**.067**	.063
		200	**.151**	**.076**	**.064**
		500	**.285**	.060	.048
		1000	**.506**	**.088**	.055

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. Boldface cells indicate a significant inflation. The Type I error rates were calculated for Indicator 1.

Table A11: The relative bias of each method across all condition with a categorical V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
ln(0.5)	ln(0.15/0.3)	100	0.868	-68.738	0.066	10.886	1.037	4.201
		200	-0.103	-70.609	0.188	5.274	1.095	-1.100
		500	-0.655	-70.100	-0.093	7.187	0.296	0.019
		1000	-0.765	-69.776	-0.182	7.636	0.361	0.684
	0	100	2.641	-58.143	1.373	25.330	2.057	2.841
		200	-0.052	-57.377	0.362	26.409	0.475	2.502
		500	-0.824	-57.515	-0.134	24.730	-0.141	0.852
		1000	-0.920	-57.805	-0.151	24.002	-0.101	0.179
	ln(0.6/0.3)	100	4.001	-40.961	2.881	49.767	2.543	5.449
		200	0.384	-41.719	0.090	50.424	-0.170	5.100
		500	-0.061	-42.833	0.364	44.790	-0.141	0.030
		1000	-0.374	-43.515	0.484	43.372	0.079	-1.288
0	ln(0.15/0.3)	100	2.379	-4.801	0.206	15.917	0.404	2.686
		200	1.894	-7.450	0.864	12.331	0.877	-0.422
		500	1.357	-5.589	0.868	13.769	0.902	0.441
		1000	0.596	-5.933	0.172	13.234	0.032	-0.099
	0	100	3.354	7.627	1.552	30.115	2.269	4.395
		200	1.823	0.293	0.850	21.750	0.880	-2.345
		500	-0.048	4.465	-0.551	26.345	-0.286	1.698
		1000	0.399	4.558	-0.008	26.491	0.138	1.013
	ln(0.6/0.3)	100	3.681	17.723	0.678	39.848	1.175	-0.492
		200	2.723	13.176	0.352	40.294	0.409	0.387
		500	0.824	14.032	-0.108	38.154	0.079	-1.273
		1000	0.559	16.281	0.034	40.892	0.195	0.458
ln(1.5)	ln(0.15/0.3)	100	3.929	30.267	0.586	17.090	0.426	0.691
		200	2.271	29.402	0.724	13.762	0.735	-1.968
		500	0.978	30.060	0.580	15.888	0.238	-0.266
		1000	1.147	29.527	0.479	15.714	0.285	-0.063
	0	100	3.487	38.315	1.172	24.858	1.244	-1.078
		200	1.298	39.273	0.355	27.437	0.250	0.755
		500	0.934	39.249	-0.147	26.205	0.252	-0.088
		1000	0.693	38.862	-0.181	25.840	0.188	-0.350
	ln(0.6/0.3)	100	4.933	51.236	1.986	41.156	1.942	1.248
		200	0.601	52.377	-0.846	42.140	-0.444	2.483
		500	0.994	50.501	-0.413	38.580	0.139	-0.493
		1000	0.827	49.117	-0.354	37.879	0.007	-1.030
ln(2)	ln(0.15/0.3)	100	1.841	55.394	0.388	18.082	0.135	0.868
		200	1.026	55.021	1.075	16.104	0.805	-0.660
		500	0.921	55.738	1.071	17.343	0.781	0.083
		1000	-0.181	56.190	0.407	17.003	0.016	-0.212
	0	100	2.458	65.858	1.344	30.309	1.847	2.793
		200	0.562	60.101	0.813	22.792	0.737	-2.950
		500	-0.863	65.404	-0.590	27.281	-0.418	0.986
		1000	0.063	64.689	-0.027	27.712	0.082	0.737
	ln(0.6/0.3)	100	1.365	74.335	0.318	37.470	0.873	-1.440
		200	0.671	73.332	0.058	38.794	0.387	0.212
		500	-0.789	71.438	-0.453	37.128	0.003	-0.917
		1000	-0.737	76.014	-0.304	39.122	0.140	0.336

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The relative bias was calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A12: The RMSE of each method across all condition with a categorical V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
ln(0.5)	ln(0.15/0.3)	100	0.130	0.215	0.129	0.216	0.140	0.184
		200	0.093	0.199	0.094	0.143	0.099	0.122
		500	0.058	0.184	0.058	0.090	0.062	0.076
		1000	0.042	0.179	0.041	0.065	0.043	0.055
	0	100	0.144	0.205	0.144	0.244	0.157	0.201
		200	0.102	0.177	0.102	0.173	0.112	0.138
		500	0.061	0.157	0.062	0.117	0.067	0.086
		1000	0.043	0.151	0.043	0.091	0.047	0.059
	ln(0.6/0.3)	100	0.168	0.204	0.167	0.302	0.188	0.232
		200	0.115	0.163	0.117	0.230	0.132	0.164
		500	0.075	0.134	0.076	0.168	0.083	0.105
		1000	0.050	0.122	0.051	0.138	0.056	0.072
0	ln(0.15/0.3)	100	0.128	0.132	0.124	0.182	0.128	0.151
		200	0.091	0.092	0.090	0.126	0.092	0.104
		500	0.058	0.058	0.058	0.081	0.058	0.063
		1000	0.040	0.044	0.040	0.064	0.041	0.046
	0	100	0.144	0.144	0.142	0.209	0.150	0.163
		200	0.098	0.101	0.097	0.142	0.100	0.113
		500	0.061	0.062	0.061	0.104	0.063	0.069
		1000	0.045	0.046	0.045	0.088	0.047	0.050
	ln(0.6/0.3)	100	0.164	0.180	0.161	0.243	0.171	0.182
		200	0.115	0.122	0.115	0.183	0.121	0.126
		500	0.071	0.080	0.071	0.133	0.075	0.079
		1000	0.053	0.065	0.053	0.121	0.055	0.054
ln(1.5)	ln(0.15/0.3)	100	0.128	0.156	0.127	0.170	0.128	0.137
		200	0.093	0.119	0.092	0.117	0.093	0.093
		500	0.058	0.096	0.058	0.080	0.058	0.058
		1000	0.040	0.084	0.041	0.063	0.041	0.043
	0	100	0.141	0.174	0.140	0.185	0.141	0.143
		200	0.100	0.141	0.099	0.136	0.101	0.099
		500	0.061	0.117	0.060	0.098	0.061	0.062
		1000	0.043	0.106	0.042	0.082	0.043	0.043
	ln(0.6/0.3)	100	0.165	0.211	0.161	0.217	0.166	0.154
		200	0.113	0.175	0.112	0.170	0.117	0.111
		500	0.073	0.147	0.074	0.130	0.076	0.072
		1000	0.051	0.133	0.049	0.112	0.051	0.050
ln(2)	ln(0.15/0.3)	100	0.129	0.198	0.124	0.167	0.124	0.134
		200	0.092	0.168	0.090	0.117	0.089	0.092
		500	0.057	0.151	0.058	0.078	0.057	0.056
		1000	0.040	0.147	0.040	0.065	0.040	0.041
	0	100	0.142	0.221	0.141	0.185	0.144	0.139
		200	0.095	0.184	0.096	0.129	0.097	0.098
		500	0.061	0.175	0.060	0.097	0.061	0.060
		1000	0.045	0.168	0.044	0.086	0.045	0.044
	ln(0.6/0.3)	100	0.165	0.251	0.160	0.205	0.163	0.150
		200	0.113	0.214	0.114	0.160	0.116	0.105
		500	0.071	0.191	0.070	0.121	0.072	0.066
		1000	0.052	0.197	0.052	0.112	0.053	0.045

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The RMSE was calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A13: The coverage rates of each method across all condition with a categorical V

h	d	N	RFA/LMS		RFA/PI		MNLFA	
			Scalar	Metric	Scalar	Metric	Scalar	Metric
ln(0.5)	ln(0.15/0.3)	100	.938	.706	.944	.930	.943	.941
		200	.927	.470	.927	.939	.936	.945
		500	.933	.112	.934	.947	.942	.958
		1000	.946	.005	.945	.946	.943	.947
	0	100	.944	.797	.946	.929	.952	.938
		200	.945	.698	.945	.931	.939	.954
		500	.953	.379	.953	.902	.952	.954
		1000	.957	.099	.958	.869	.958	.958
	ln(0.6/0.3)	100	.946	.891	.949	.925	.955	.950
		200	.953	.850	.951	.909	.943	.940
		500	.940	.715	.934	.856	.938	.937
		1000	.956	.476	.955	.759	.954	.940
0	ln(0.15/0.3)	100	.946	.936	.944	.929	.950	.944
		200	.938	.941	.938	.934	.940	.937
		500	.938	.951	.942	.914	.944	.959
		1000	.950	.929	.952	.894	.954	.944
	0	100	.934	.930	.929	.918	.934	.933
		200	.956	.943	.956	.924	.959	.934
		500	.953	.964	.952	.864	.948	.954
		1000	.951	.940	.950	.769	.948	.944
	ln(0.6/0.3)	100	.952	.908	.939	.914	.935	.928
		200	.941	.928	.940	.889	.945	.947
		500	.947	.934	.951	.828	.954	.945
		1000	.933	.883	.932	.682	.937	.961
ln(1.5)	ln(0.15/0.3)	100	.939	.903	.940	.914	.942	.947
		200	.933	.864	.932	.927	.933	.948
		500	.932	.755	.936	.917	.940	.956
		1000	.946	.566	.940	.879	.940	.950
	0	100	.943	.881	.943	.921	.948	.938
		200	.943	.838	.945	.911	.938	.951
		500	.953	.653	.955	.853	.953	.959
		1000	.944	.404	.956	.782	.959	.954
	ln(0.6/0.3)	100	.952	.856	.952	.904	.952	.940
		200	.951	.771	.952	.868	.951	.949
		500	.947	.568	.938	.790	.941	.937
		1000	.952	.316	.953	.643	.954	.944
ln(2)	ln(0.15/0.3)	100	.940	.811	.941	.916	.951	.930
		200	.928	.690	.941	.913	.944	.938
		500	.937	.390	.940	.887	.944	.954
		1000	.943	.097	.950	.845	.953	.944
	0	100	.927	.758	.930	.900	.929	.935
		200	.962	.683	.957	.902	.957	.934
		500	.948	.257	.953	.835	.952	.951
		1000	.943	.056	.952	.703	.945	.942
	ln(0.6/0.3)	100	.936	.765	.940	.890	.936	.934
		200	.950	.621	.940	.865	.946	.951
		500	.948	.282	.950	.766	.954	.944
		1000	.935	.039	.932	.583	.934	.953

Note. h = the effect of V on the common-factor variance, d = the effect of V on the indicator's residual variance, N = total sample size. The coverage rates were calculated for b_2 (reflecting a violation of scalar invariance) and c_4 (reflecting a violation of metric invariance).

Table A14: The power and Type I error rates for new conditions with a categorical V

Method	P	Violations	Heterogeneity	Power		Type I error
				Scalar	Metric	
RFA/LMS	10	40%	30%	1.000	.884	**.073**
			90%	1.000	.922	**.077**
		60%	30%	1.000	.891	**.072**
			90%	1.000	.920	**.081**
	20	40%	30%	.996	.919	**.083**
			90%	.997	.936	**.095**
		60%	30%	.996	.911	**.085**
			90%	.994	.928	**.100**
RFA/PI	10	40%	30%	1.000	.649	**.064**
			90%	1.000	.778	**.070**
		60%	30%	1.000	.661	**.064**
			90%	1.000	.788	**.068**
	20	40%	30%	1.000	.709	**.072**
			90%	1.000	.801	**.069**
		60%	30%	1.000	.728	**.070**
			90%	1.000	.787	**.069**
MNLFA	10	40%	30%	1.000	.658	.060
			90%	1.000	.720	**.070**
		60%	30%	1.000	.674	.059
			90%	1.000	.729	**.071**
	20	40%	30%	1.000	.696	**.066**
			90%	1.000	.753	**.065**
		60%	30%	1.000	.693	.063
			90%	1.000	.745	**.065**

Note. P = the total number of indicators, violations = the percentage of indicators that violate measurement invariance, heterogeneity = the percentage of indicators with unequal residual variances with respect to V. Boldface cells indicate a significant inflation. Power was calculated for Indicator 2 (violating scalar invariance, that is, $b_2 \neq 0$) and Indicator 4 (violating metric invariance, that is, $c_4 \neq 0$). The Type I error rates were calculated for Indicator 1.

Appendix III: Supplementary Figures
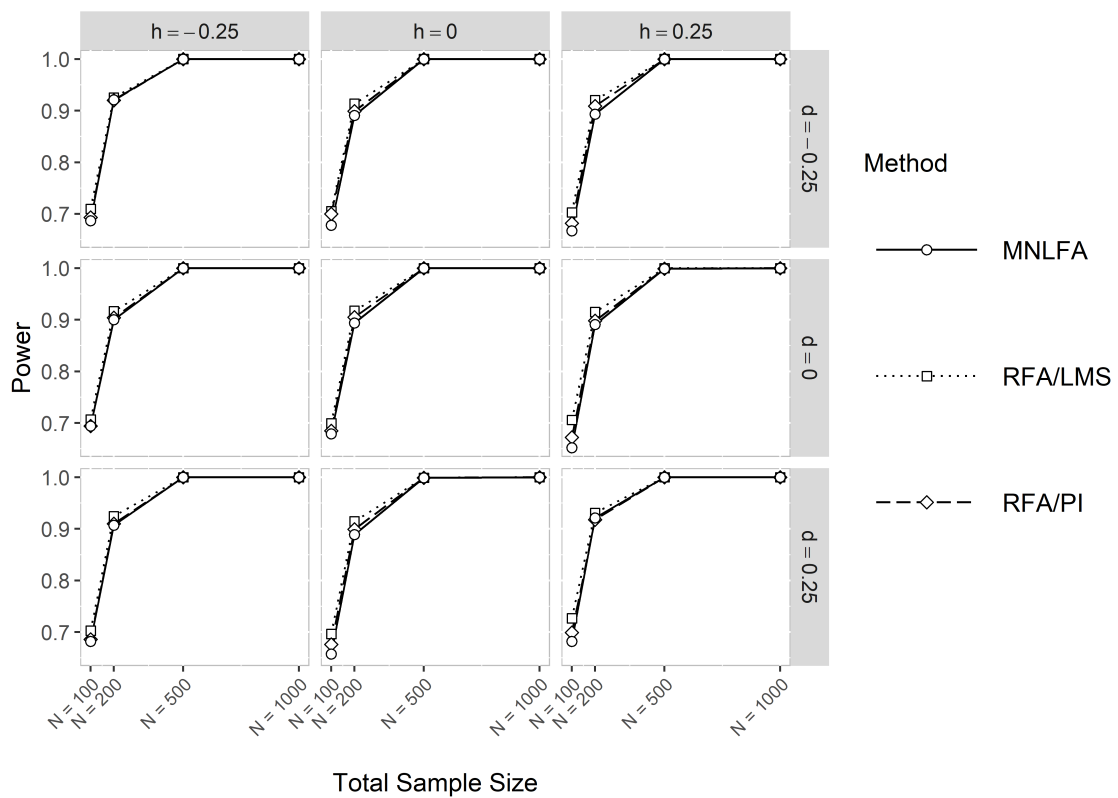
Figure A1: The power to detect a violation of scalar invariance of Indicator 2 (i.e., $b_2 \neq 0$) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
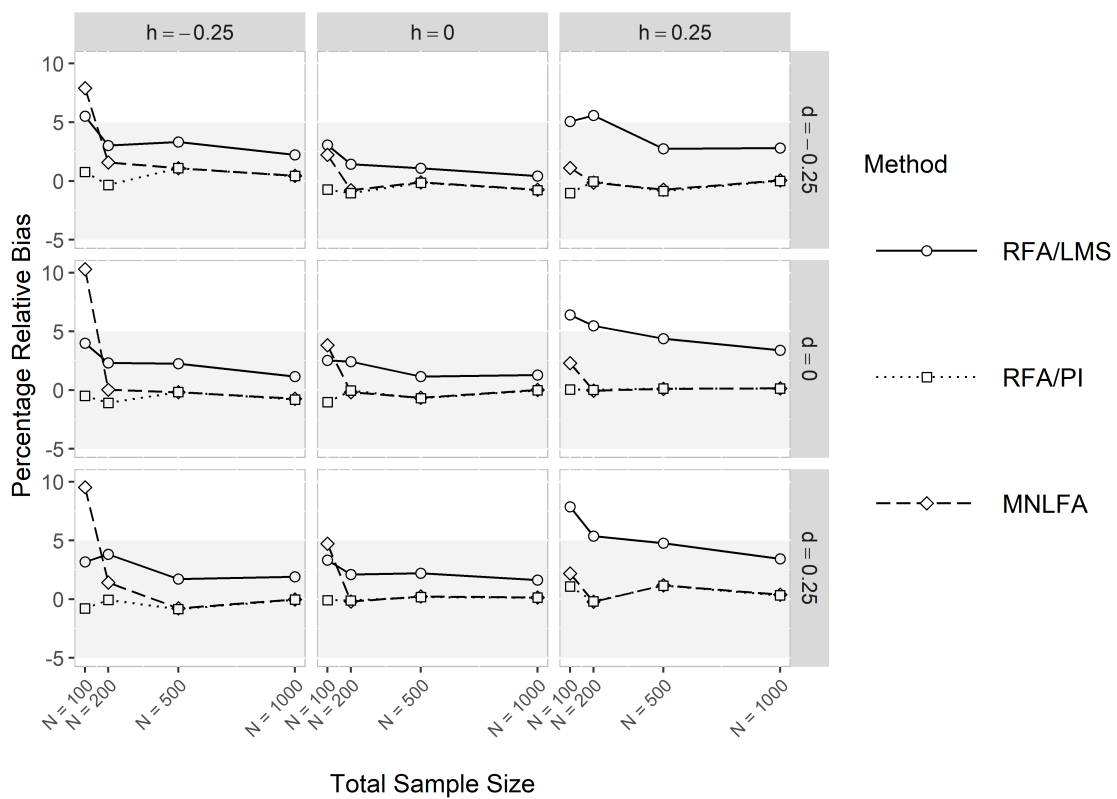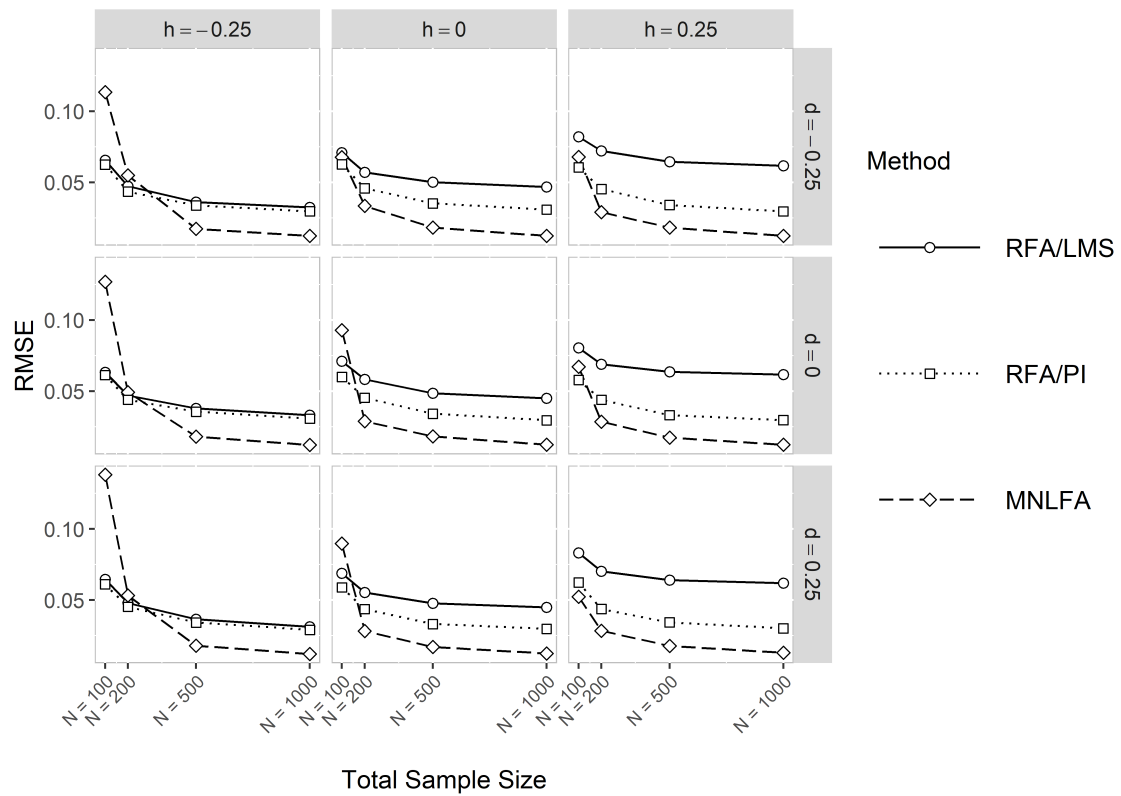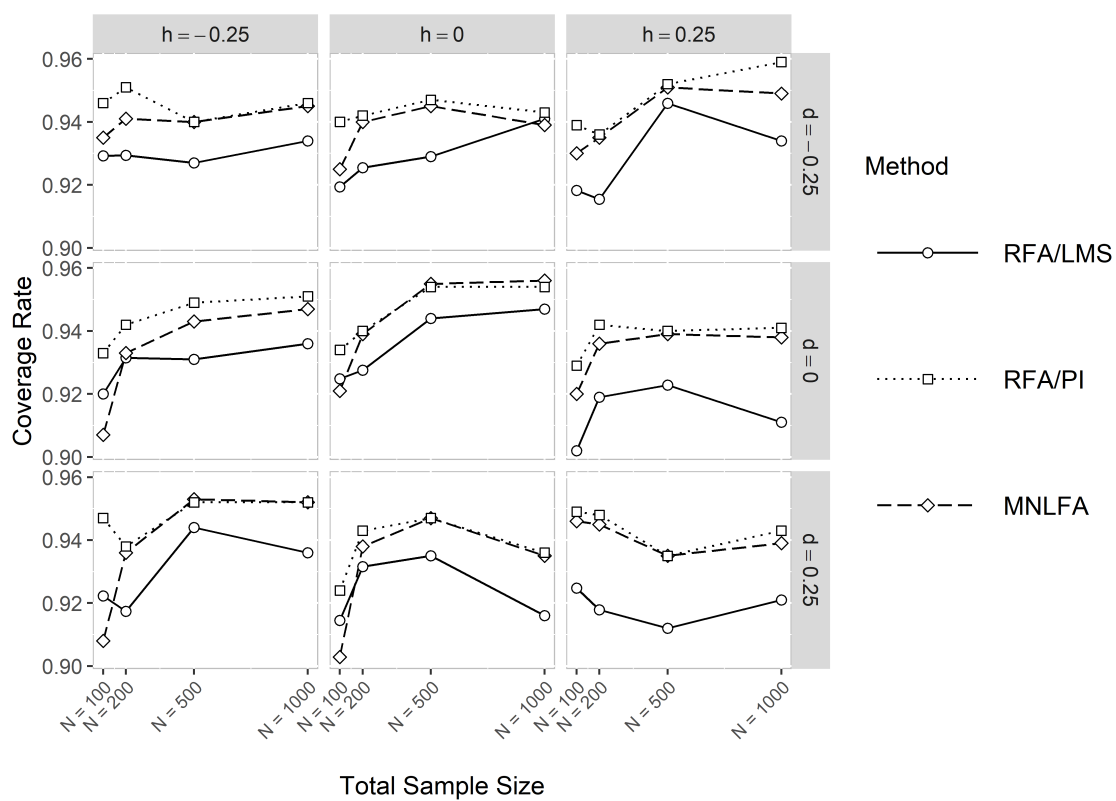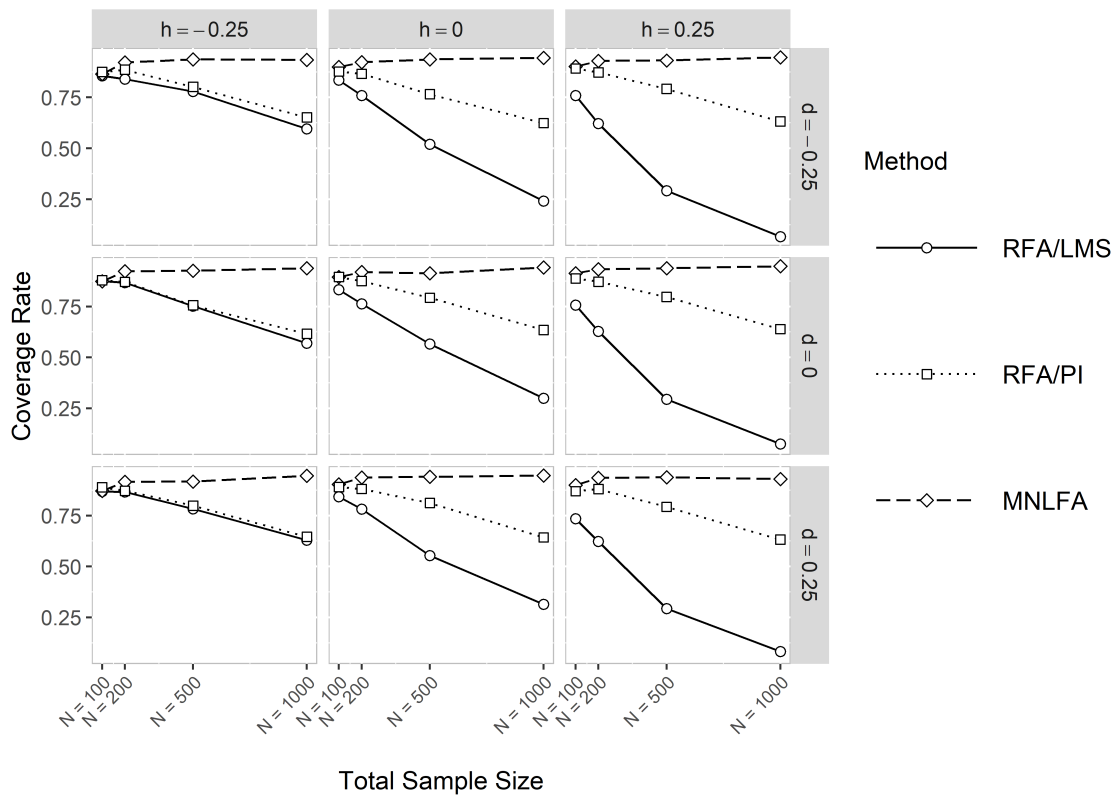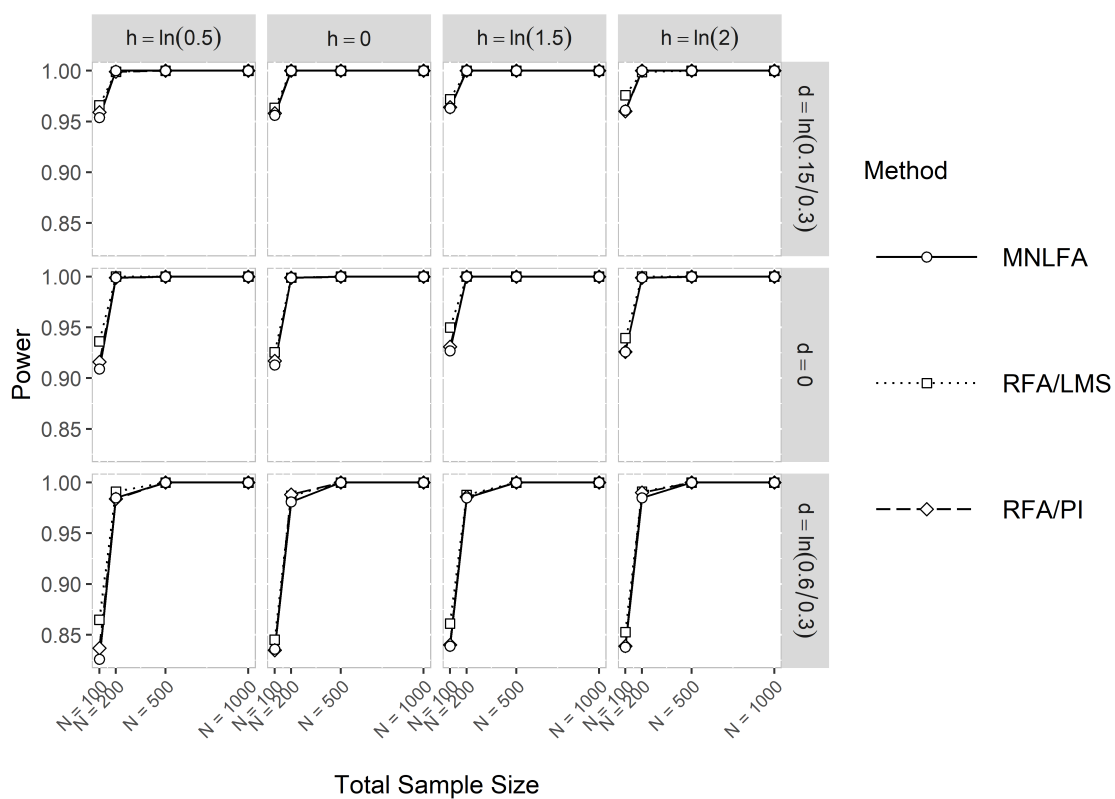
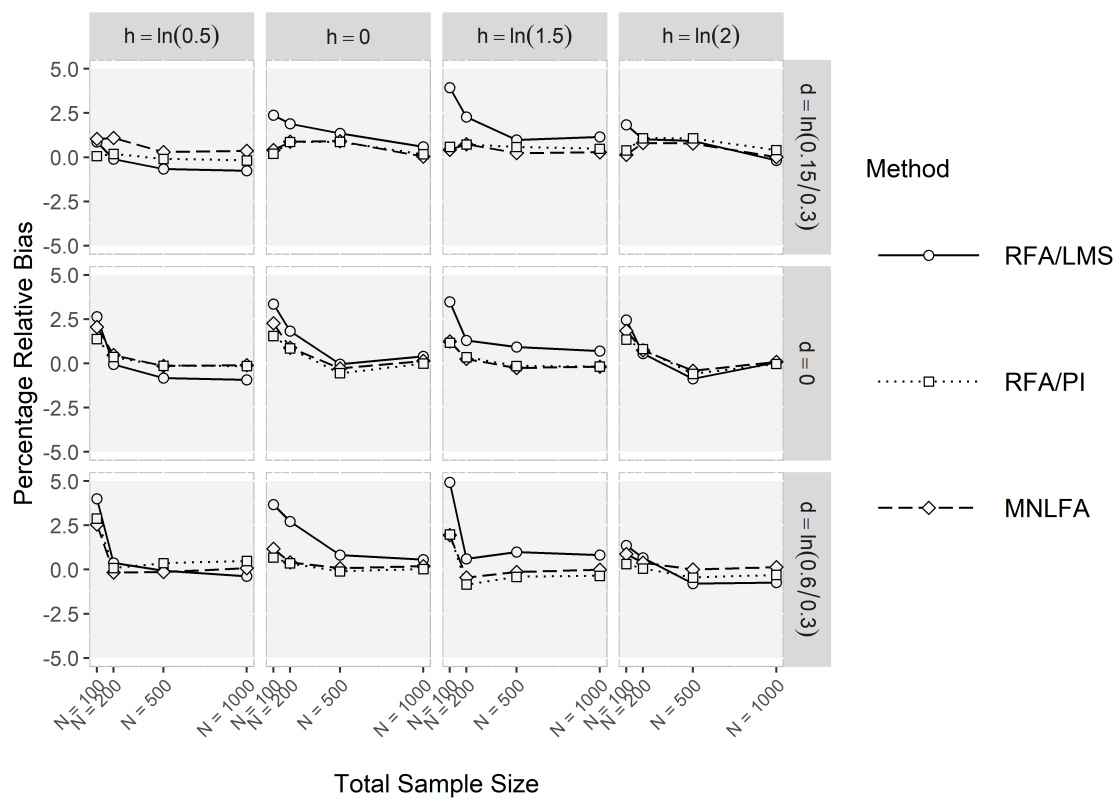Figure A2: The relative bias of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
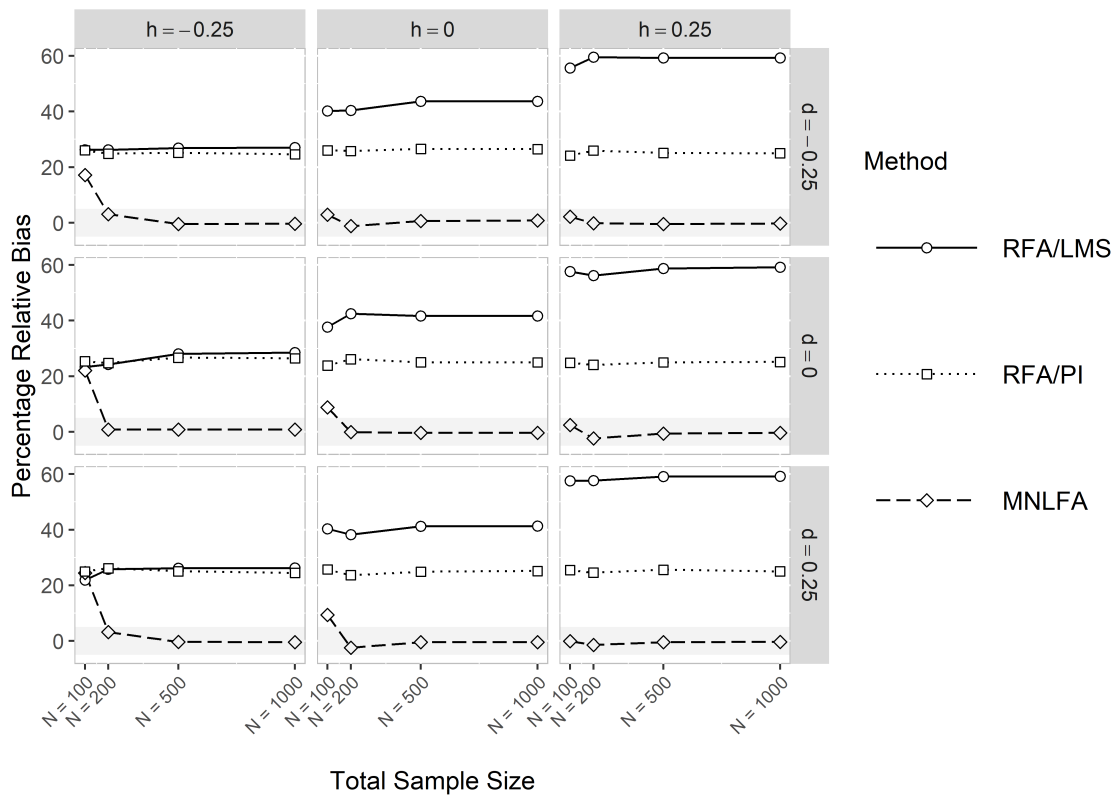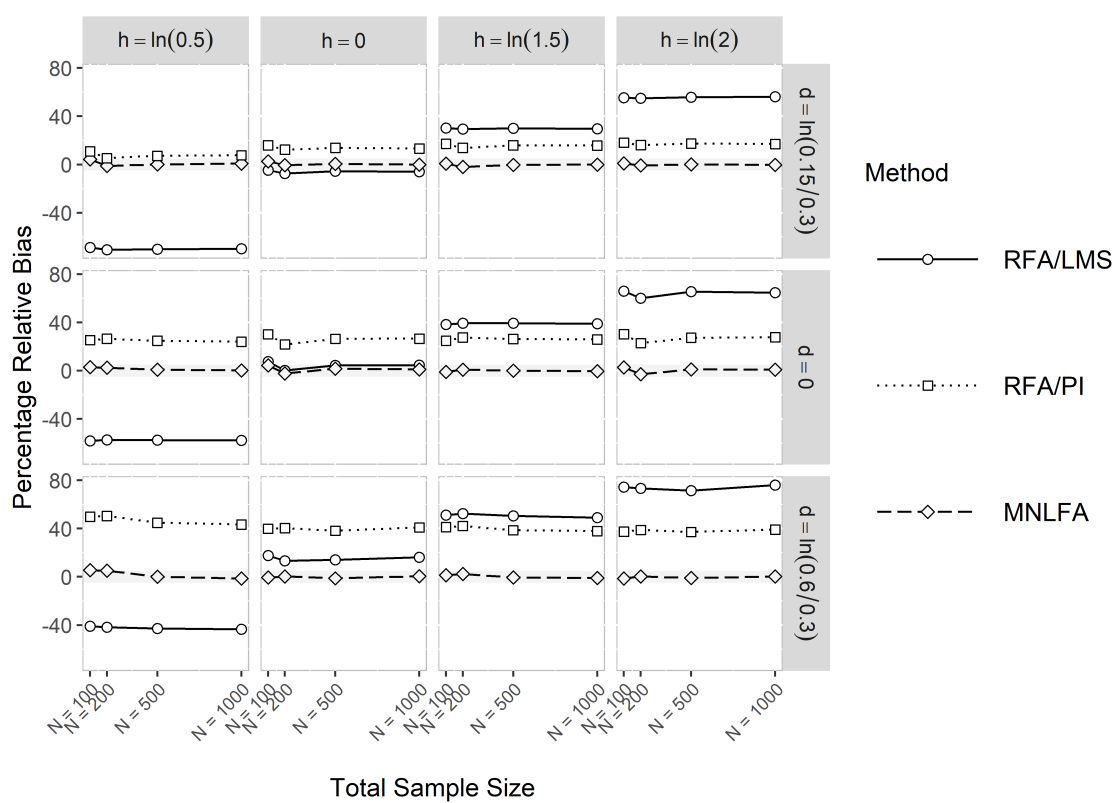
Figure A3: The relative bias of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
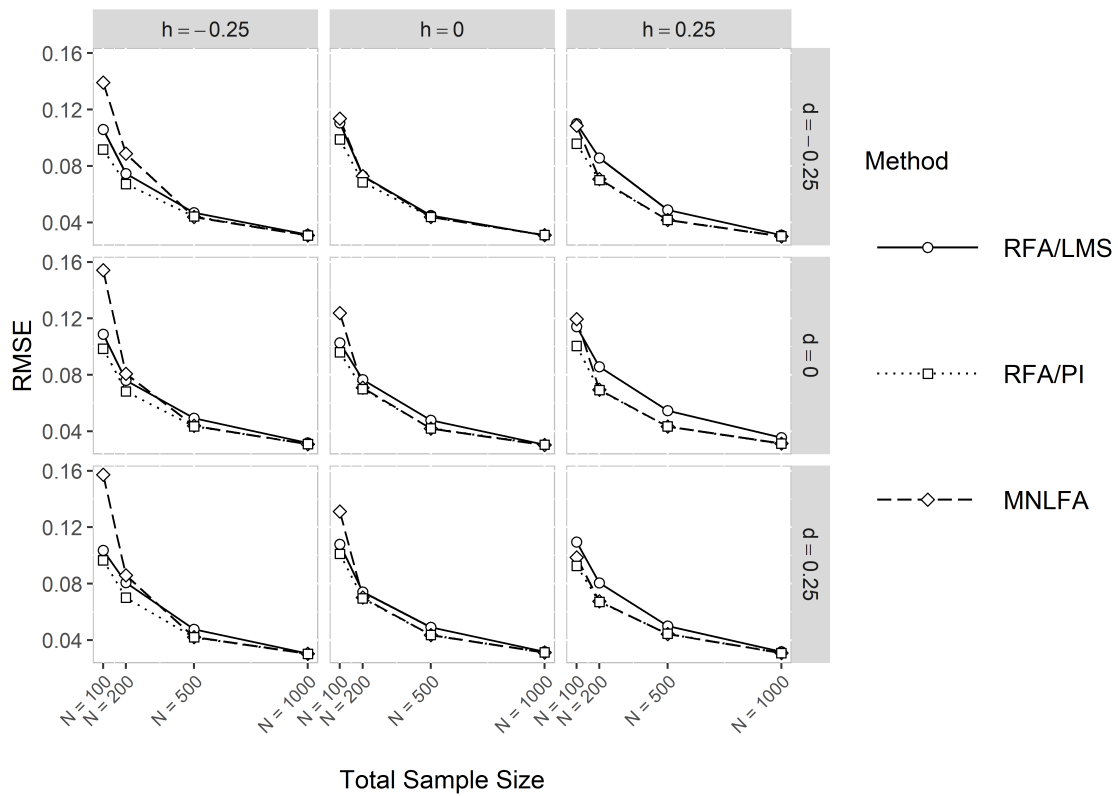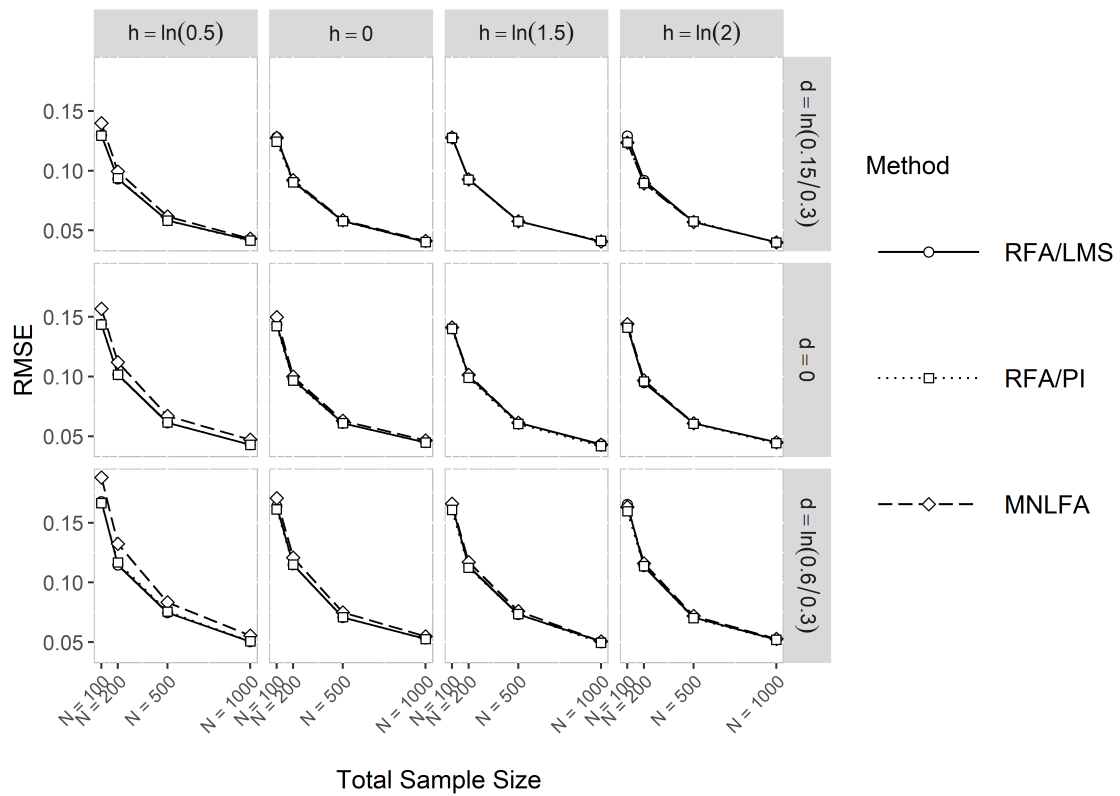
Figure A4: The RMSE of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A5: The RMSE of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A6: The coverage rates of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
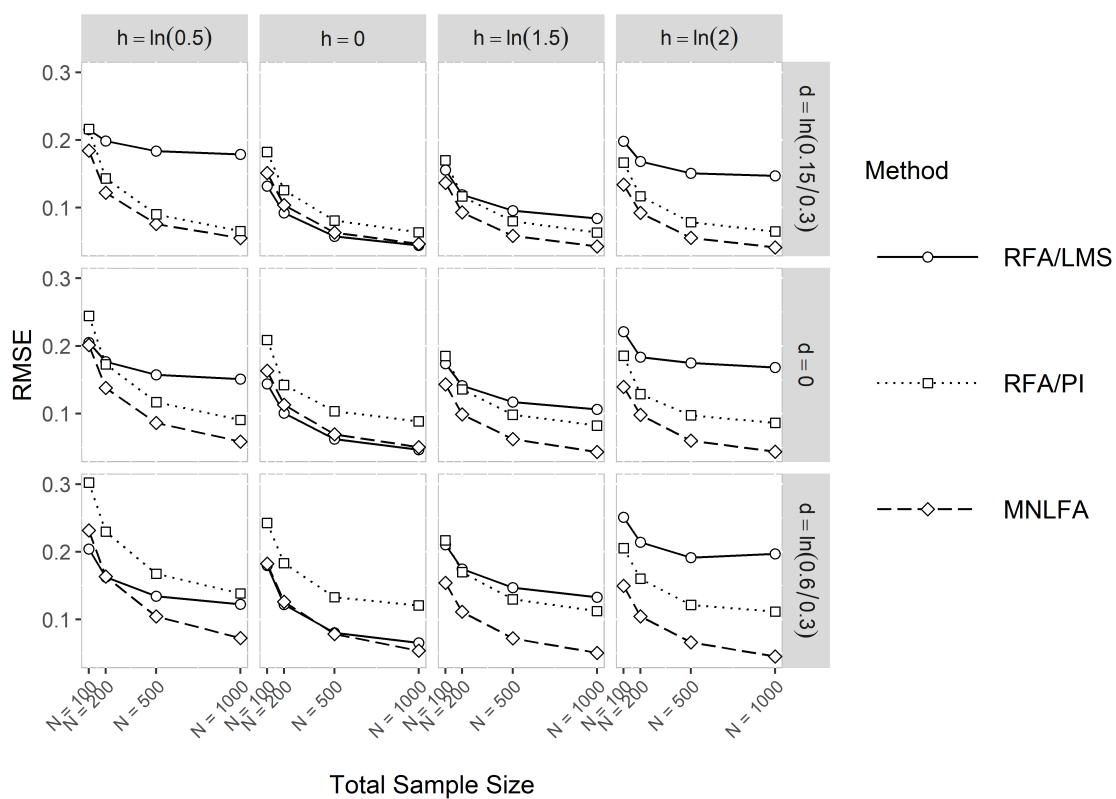
Figure A7: The coverage rates of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a continuous V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A8: The power to detect a violation of scalar invariance of Indicator 2 (i.e., $b_2 \neq 0$) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
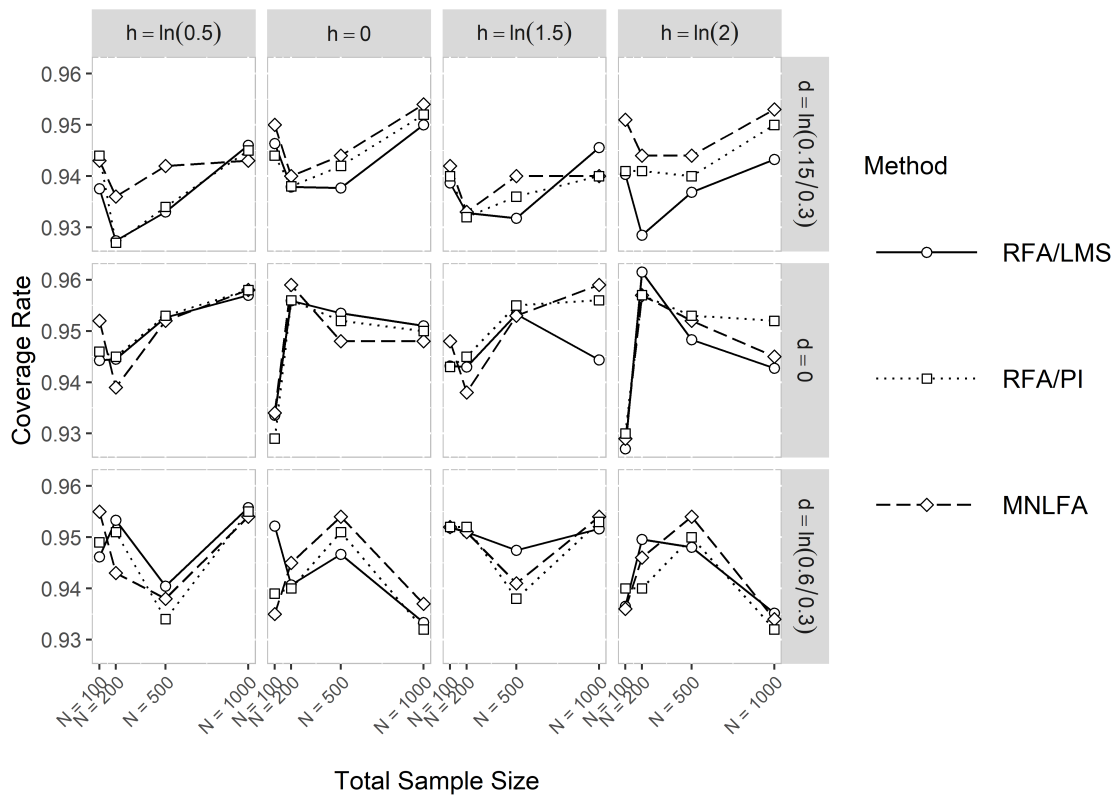
Figure A9: The relative bias of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.
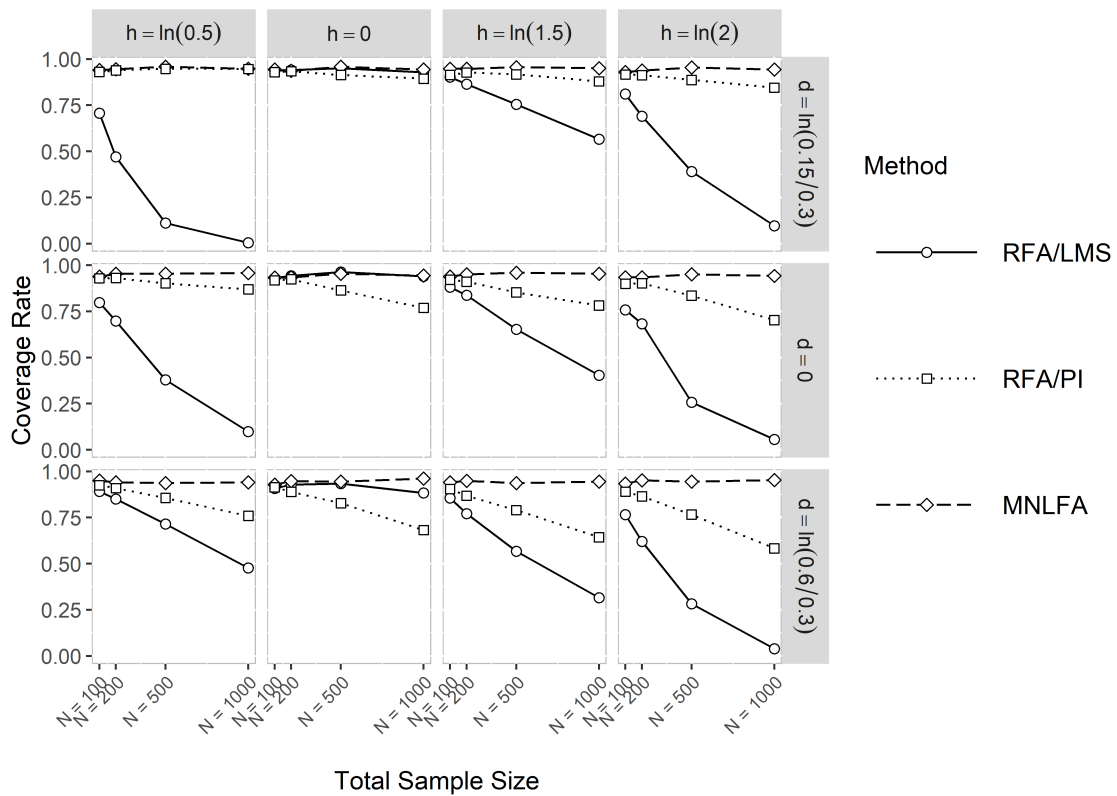
Figure A10: The relative bias of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A11: The RMSE of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A12: The RMSE of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A13: The coverage rates of the parameter estimate b_2 (i.e., a violation of scalar invariance of Indicator 2) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

Figure A14: The coverage rates of the parameter estimate c_4 (i.e., a violation of metric invariance of Indicator 4) across all conditions with a categorical V. Note that h is the effect of V on the common-factor variance, and d is the effect of V on the indicator's residual variance.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, *52*(2), 119–126. doi: 10.1080/00031305.1998.10480550

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). New York, NY: Springer New York. doi: 10.1007/978-1-4612-1694-0_15

Barendse, M. T., Oort, F. J., & Garst, G. J. A. (2010). Using restricted factor analysis with latent moderated structures to detect uniform and nonuniform measurement bias: A simulation study. *Advances in Statistical Analysis*, *94*, 117–127. doi: 10.1007/s10182-010-0126-1

Barendse, M. T., Oort, F. J., Werner, C. S., Ligtvoet, R., & Schermelleh-Engel, K. (2012). Measurement bias detection through factor analysis. *Structural Equation Modeling*, *19*(4), 561–579. doi: 10.1080/10705511.2012.713261

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

Bauer, D. J., Belzak, W. C., & 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*, *27*(1), 43–55. doi: 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. doi: 10.1037/a0015583

Belzak, W. C. M. (2021). `regDIF`: Regularized differential item functioning [Computer software manual]. Pittsburgh, PA. Retrieved from `https://github.com/wbelzak/regDIF/` (version 1.0.0)

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., ... others (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika*, *76*(2), 306–317. doi: 10.1007/S11336-010-9200-6

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

Brennan, R. L. (2004). Revolutions and evolutions in current educational testing. *The Iowa Academy of Education, Occasional Research Paper*.

Buse, A. (1982). The likelihood ratio, wald, and lagrange multiplier tests: An expository note. *The American Statistician*, *36*(3), 153–157. doi: 10.2307/2683166

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi: 10.1037/0033-2909.105.3.456

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied psychological measurement*, *12*(3), 253–260. doi: 10.1177/014662168801200304

Chalmers, R. P. (2012). `mirt`: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi: 10.18637/jss.v048.i06

Cheng, C. H., & Watkins, D. (2000). Age and gender invariance of self-concept factor structure: An investigation of a newly developed chinese self-concept instrument. *International Journal of Psychology*, *35*(5), 186–193. doi: 10.1080/00207590050171120

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. doi: 10.1177/014920639902500101

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, *40*(7), 486–499. doi: 10.1177/0146621616659738

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16–29. doi: 10.1037/1082-989X.1.1.16

de Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment*, *17*(2), 168–178. doi: 10.1037/1040-3590.17.2.168

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Deng, L., & Yuan, K.-H. (2016). Comparing latent means without mean structure models: A projection-based approach. *Psychometrika*, *81*(3), 802–829. doi: 10.1007/s11336-015-9491-8

References

Denollet, J. (2005). DS14: Standard assessment of negative affectivity, social inhibition, and Type D personality. *Psychosomatic Medicine*, *67*(1), 89–97. doi: 10.1097/01.psy.0000149256.81953.49

Denollet, J., Pedersen, S. S., Vrints, C. J., & Conraads, V. M. (2013). Predictive value of social inhibition and negative affectivity for cardiovascular events and mortality in patients with coronary artery disease: The type D personality construct. *Psychosomatic Medicine*, *75*(9), 873–881. doi: 10.1097/PSY.0000000000000001

Dimitruk, P., Schermelleh-Engel, K., Kelava, A., & Moosbrugger, H. (2007). Challenges in nonlinear structural equation modeling. *Methodology*, *3*(3), 100–114. doi: 10.1027/1614-2241.3.3.100

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin*, *95*(1), 134–135. doi: 10.1037/0033-2909.95.1.134

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*(4), 662–680. doi: 10.1037/0021-9010.70.4.662

Finch, H. W., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling*, *25*(5), 673–686. doi: 10.1080/10705511.2018.1431781

Finch, H. W., French, B. F., & Hernández Finch, M. E. (2018). Comparison of methods for factor invariance testing of a 1-factor model with small samples and skewed latent traits. *Frontiers in Psychology*, *9*, 1–12. doi: 10.3389/fpsyg.2018.00332

Finney, S. J., & DiStefano, C. (2006). Non-normal and categorical data in structural equation modeling. In G. Hancock & R. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439–492). IAP Information Age Publishing.

Goodrich, S., & Ercikan, K. (2019). Measurement comparability of reading in the english and french canadian populations: Special case of the 2011 progress in international reading literacy study. *Frontiers in Education*, *4*, 1–15. doi: 10.3389/feduc.2019.00120

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*(4), 621–638. doi: 10.1080/10705511.2017.1402334

Harpole, J. K. (2015). *A Bayesian MIMIC model for testing non-uniform DIF in two and three groups* (Doctoral dissertation, University of Kansas). Retrieved from `http://hdl.handle.net/1808/21697`

Henseler, J., & Chin, W. W. (2010). A comparison of approaches for the analysis of interaction effects between latent variables using partial least squares path modeling. *Structural Equation Modeling*, *17*(1), 82–109. doi: 10.1080/10705510903439003

Hidalgo-Montesinos, M. D., & Lopez-Pina, J. A. (2002). Two-stage equating in differential item functioning detection under the graded response model with the raju area measures and the lord statistic. *Educational and Psychological Measurement*, *62*(1), 32–44. doi: 10.1177/0013164402062001003

Hildebrandt, A., Lüdtke, O., Robitzsch, A., Sommer, C., & Wilhelm, O. (2016). Exploring factor model parameters across continuous variables with local structural equation models. *Multivariate Behavioral Research*, *51*(2–3), 257–258. doi: 10.1080/00273171.2016.1142856

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi: 10.1080/03610739208253916

Jiang, G., Mai, Y., & Yuan, K.-H. (2017). Advances in measurement invariance and mean comparison of latent variables: Equivalence testing and a projection-based approach. *Frontiers in Psychology*, *8*, 1–13. doi: 10.3389/fpsyg.2017.01823

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

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, *70*(351a), 631–639. doi: 10.1080/01621459.1975.10482485

Jöreskog, K. G., & Yang-Wallentin, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced Structural Equation Modeling: Issues and Techniques* (pp. 57–88). Mahwah, NJ: Lawrence Erlbaum Associates.

Jorgensen, T. D. (2017). Applying permutation tests and multivariate modification indices to configurally invariant models that need respecification. *Frontiers in Psychology*, *8*, 1–9. doi: 10.3389/fpsyg.2017.01455

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A., & Rosseel, Y. (2018). `semTools`: Useful tools for structural equation modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=semTools` (R package version 0.5-0)

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A., & Rosseel, Y. (2019). `semTools`: Useful tools for structural equation modeling [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=semTools` (R package version 0.5-1.935)

References

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*(1), 201–210. doi: 10.1037/0033-2909.96.1.201

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika*, *65*(4), 457–474. doi: 10.1007/BF02296338

Klein, A. G., & Muthén, B. O. (2007). Quasi-maximum likelihood estimation of structural equation models with multiple interaction and quadratic effects. *Multivariate Behavioral Research*, *42*(4), 647–673. doi: 10.1080/00273170701710205

Kline, R. (2011). *Principles and practice of structural equation modeling (3rd ed.)*. New York, NY: Guilford.

Kolbe, L., & Jorgensen, T. D. (2018). Using product indicators in restricted factor analysis models to detect nonuniform measurement bias. In M. Wiberg, S. A. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology: The 82nd Annual Meeting of the Psychometric Society, Zurich, Switzerland, 2017* (pp. 235–245). New York, NY: Springer. doi: 10.1007/978-3-319-77249-3_20

Kolbe, L., & Jorgensen, T. D. (2019). Using restricted factor analysis to select anchor items and detect differential item functioning. *Behavior Research Methods*, *51*, 138–151. doi: 10.3758/s13428-018-1151-3

Kolbe, L., Jorgensen, T. D., & Molenaar, D. (2021). The impact of unmodeled heteroskedasticity on assessing measurement invariance in single-group models. *Structural Equation Modeling*, *28*(1), 82–98. doi: 10.1080/10705511.2020.1766357

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56. doi: 10.1177/0013164414529792

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103. doi: 10.1177/0146621614544195

Lin, G.-C., Wen, Z., Marsh, H. W., & Lin, H.-S. (2010). Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling*, *17*(3), 374–391. doi: 10.1080/10705511.2010.488999

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*(1), 53–76. doi: 10.1207/s15327906mbr3201_3

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling*, *13*(4), 497–519. doi: 10.1207/s15328007sem1304_1

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486–506. doi: 10.1037/met0000075

Lodder, P., Denollet, J., Emons, W. H., Nefs, G., Pouwer, F., Speight, J., & Wicherts, J. M. (2019). Modeling interactions between latent variables in research on type d personality: A monte carlo simulation and clinical study of depression and anxiety. *Multivariate Behavioral Research*, *54*(5), 637–665. doi: 10.1080/00273171.2018.1562863

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. doi: 10.1037/1082-989X.7.1.19

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. doi: 10.3758/BRM.42.3.847

Marsh, H. W. (1994). Confirmatory factor analysis models of factorial invariance: A multifaceted approach. *Structural Equation Modeling*, *1*(1), 5–34. doi: 10.1080/10705519409539960

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: The extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods*, *23*(3), 524–545. doi: 10.1037/met0000113

Marsh, H. W., Wen, Z., & Hau, K.-T. (2004). Structural equation models of latent interactions: Evaluation of alternative estimation strategies and indicator construction. *Psychological Methods*, *9*(3), 275–300. doi: 10.1037/1082-989X.9.3.275

Masyn, K. E. (2017). Measurement invariance and differential item functioning in latent class analysis with stepwise multiple indicator multiple cause modeling. *Structural Equation Modeling*, *24*(2), 180–197. doi: 10.1080/10705511.2016.1254049

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. doi: 10.1037/1082-989X.10.3.259

Mehta, P. D., Neale, M. C., & Flay, B. R. (2004). Squeezing interval change from ordinal panel data: Latent growth curves with ordinal outcomes. *Psychological Methods*, *9*(3), 301–333. doi: 10.1037/1082-989X.9.3.301

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

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research*, *29*(3), 223–236. doi: 10.1207/s15327906mbr2903_2

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69–S77. doi: 10.1097/01.mlr.0000245438.73837.89

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584. doi: 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. doi: 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. doi: 10.21500/20112084.857

Millsap, R. E., & Tein, J.-Y. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi: 10.1207/S15327906MBR3903_4

Molenaar, D. (2015). Heteroscedastic latent trait models for dichotomous data. *Psychometrika*, *80*(3), 625–644. doi: 10.1007/s11336-014-9406-0

Molenaar, D. (2020). A flexible moderated factor analysis approach to test for measurement invariance across a continuous variable. *Psychological Methods*, *26*(6). doi: 10.1037/met0000360

Molenaar, D., Dolan, C. V., & De Boeck, P. (2012). The heteroscedastic graded response model with a skewed latent trait: Testing statistical and substantive hypotheses related to skewed item category functions. *Psychometrika*, *77*(3), 455–478. doi: 10.1007/S11336-012-9273-5

Molenaar, D., Dolan, C. V., & van der Maas, H. L. J. (2011). Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling*, *18*(4), 578–594. doi: 10.1080/10705511.2011.607095

Molenaar, D., Dolan, C. V., & Verhelst, N. D. (2010). Testing and modelling non-normality within the one-factor model. *British Journal of Mathematical and Statistical Psychology*, *63*(2), 293–317. doi: 10.1348/000711009X456935

Molenaar, D., Dolan, C. V., Wicherts, J. M., & van der Maas, H. L. J. (2010). Modeling differentiation of cognitive abilities within the higher-order factor model using moderated factor analysis. *Intelligence*, *38*(6), 611–624. doi: 10.1016/j.intell.2010.09.002

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. doi: 10.1007/BF02294210

Muthén, B., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in mplus. *Mplus web notes*, *4*(5), 1–22.

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: Alignment and random effects. *Sociological Methods & Research*, *47*(4), 637–664. doi: 10.1177/0049124117701488

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

Muthén, L. K., & Muthén, B. O. (2012). M*plus* user's guide (7th ed.) [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Neale, M. C. (1998). Modeling interaction and nonlinear effects with Mx: A general approach. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Interaction and Non-linear Effects in Structural Equation Modeling* (pp. 43–61). Lawrence Erlbaum Associates Publishers.

Neale, M. C., Aggen, S. H., Maes, H. H., Kubarych, T. S., & Schmitt, J. E. (2006). Methodological issues in the assessment of substance use phenotypes. *Addictive Behaviors*, *31*(6), 1010–1034. doi: 10.1016/j.addbeh.2006.03.047

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., . . . Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535-549. doi: 10.1007/s11336-014-9435-8

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika*, *6*(2), 150–166.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*(2), 107–124. doi: 10.1080/10705519809540095

Purcell, S. (2002). Variance components models for gene–environment interaction in twin analysis. *Twin Research and Human Genetics*, *5*(6), 554–571. doi: 10.1375/twin.5.6.554

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. doi: 10.1016/j.dr.2016.06.004

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

R Core Team. (2018). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. doi: 10.1037/a0029315

Robitzsch, A. (2019). `mnlfa`: Moderated nonlinear factor analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=mnlfa` (R package version 0.1-53)

Robitzsch, A. (2020a). `sirt`: Supplementary item response theory models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=sirt` (R package version 3.9-4)

Robitzsch, A. (2020b). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*, 1–7. doi: 10.3389/feduc.2020.589965

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. doi: 10.1007/BF02296192

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika*, *75*(2), 243–248. doi: 10.1007/s11336-009-9135-y

Schauberger, G. (2021). `GPCMlasso`: Differential item functioning in generalized partial credit models [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=GPCMlasso` (R package version 0.1-5)

Schulze, D., & Pohl, S. (2021). Finding clusters of measurement invariant items for continuous covariates. *Structural Equation Modeling*, *28*(2), 219–228. doi: 10.1080/10705511.2020.1771186

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

Stan Development Team. (2021). `RStan`: The R interface to Stan [Computer software manual]. Retrieved from `https://mc-stan.org/` (R package version 2.21.3)

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–1306. doi: 10.1037/ 0021-9010.91.6.1292

Steenkamp, J.-B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–90. doi: 10.1086/209528

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. doi: 10.1007/S11336-013-9388-3

Sudarshan, N. J., Bowden, S. C., Saklofske, D. H., & Weiss, L. G. (2016). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale–IV. *Psychological Assessment*, *28*(11), 1489–1501. doi: 10.1037/pas0000290

Suh, Y. (2015). The performance of maximum likelihood and weighted least square mean and variance adjusted estimators in testing differential item functioning with nonnormal trait distributions. *Structural Equation Modeling*, *22*(4), 568–580. doi: 10.1080/10705511.2014.937669

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408. doi: 10.1007/ BF02294363

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727–750. doi: 10.1007/s11336-015 -9488-3

Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package nlsem. *Journal of Statistical Software*, *77*(7), 1–20. doi: 10.18637/jss.v077.i07

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*, *24*(4), 585–598. doi: 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*, 1–22. doi: 10.1007/s00180-018-0815-x

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19. Retrieved from `https://www.jstatsoft.org/article/view/v020i11`

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. doi: 10.1177/109442810031002

Van de Vijver, F., & Fischer, R. (2009). Improving methodological robustness in cross-cultural organizational research. *Handbook of culture, organizations, and work*, 491–517.

Wang, M., & Woods, C. M. (2017). Anchor selection using the wald test anchor-all-test-all procedure. *Applied Psychological Measurement*, *41*(1), 17–29. doi: https://doi.org/10.1177/0146621616668014

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of rasch models. *The Journal of Experimental Education*, *72*(3), 221–261. doi: 10.3200/JEXE.72.3.221-261

Wirth, R., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58–79. doi: 10.1037/1082-989X.12.1.58

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of nonnormality. *Applied Psychological Measurement*, *32*(7), 511–526. doi: 10.1177/0146621607310402

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. doi: 10.1177/0146621607314044

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339–361. doi: 10.1177/0146621611405984

Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*(4), 1014–1045. doi: 10.1007/s11336-016-9506-0

Summary

Novel Approaches to Assess Measurement Invariance

The measurement of latent constructs like mathematical ability or social anxiety serves an important role in social and behavioral sciences. Because such constructs cannot be measured directly, observed measures function as indicators of the latent construct. In order to meaningfully compare a latent construct across individuals or groups, the relationship between each observed indicator and the latent construct should be the same for each individual or group. This condition is also referred to as measurement invariance. In other words, if measurement invariance with respect to a specific background variable holds, the measurement of the latent construct is invariant across that background variable. If measurement invariance does not hold, differences in observed scores across individuals or groups may arise from differences on the background variable instead of differences on the latent construct. It is therefore important to assess measurement invariance before comparing individuals or groups on latent constructs.

A common class of methods to assess measurement invariance within the structural equation modeling (SEM) framework is confirmatory factor analysis (CFA). One of the traditional CFA methods to evaluate measurement invariance across a categorical background variable is multiple-group CFA (MGCFA; Vandenberg & Lance, 2000). In MGCFA, a CFA model is estimated for each group and measurement invariance is assessed by comparing the fit of models with and without equality constraints on the measurement parameters across the background variable. In addition to MGCFA, single-group methods have been proposed for the purpose of assessing measurement invariance, including restricted factor analysis (RFA; Oort, 1992) and moderated non-linear factor analysis (MNLFA; Bauer & Hussong, 2009). These single-group methods involve fitting a single CFA model to the data aggregated over the background variable and are therefore more suitable for smaller samples sizes and for testing for measurement invariance across multiple continuous and categorical background variables simultaneously.

The current dissertation focuses on the performance of novel ways of assessing measurement invariance using single-group methods. The dissertation starts with a study of the single-group method RFA, which is readily suited to assess scalar invariance, but requires an extensive method to evaluate metric invariance. RFA is most commonly extended with latent moderated structural equations (LMS). Although LMS has shown to obtain high power to detect violations of metric invariance, severely inflated Type I error rates have also been observed when using this method in RFA models (see Barendse et al., 2010, 2012; Woods & Grimm, 2011). Therefore, we propose PI as an alternative to LMS in RFA models in **Chapter 2**. Using a single simulated dataset, we show how the PI method can be used in RFA models to assess metric invariance. We find results that are comparable to the results obtained with LMS, which indicates that the PI method is a viable alternative.

The performance of PI in RFA models is investigated more extensively with a simulation study presented in **Chapter 3**. In specific, we evaluate its Type I error rates and power to detect violations of scalar and metric invariance in comparison to the more

traditional LMS method. The results of the simulation study show that the PI method obtains similar power but lower Type I error rates compared to LMS in almost all simulation conditions. The Type I error rates observed in conditions with PI are all close to the nominal level of significance. This indicates that using PI in RFA models can minimize the probability of false conclusions regarding which observed indicators violate the assumption of scalar or metric invariance. In line with previous studies (see Barendse et al., 2010, 2012; Woods & Grimm, 2011), severely inflated Type I error rates are observed in conditions with LMS.

A possible explanation for the inflated Type I error rates observed with LMS in RFA models is a violation of the assumption of homoskedasticity (i.e., equal common-factor and residual variances; Chun et al., 2016; Meredith & Teresi, 2006). In order to confirm or disconfirm this possible explanation, we investigate the impact of violations of this assumption on the performance of RFA combined with LMS and PI in **Chapter 4**. In contrast to RFA, the recently proposed MNLFA (Bauer & Hussong, 2009; Bauer, 2017) method for assessing measurement invariance does not require assuming homoskedasticity with respect to the background variable. Hence, we also include a comparison between RFA and MNLFA under each of the different simulation conditions. The results of the simulation study presented in this chapter show that the Type I error rates obtained by RFA/LMS substantially increase as a function of heteroskedasticity (i.e., unequal common-factor and residual variances), whereas MNLFA and RFA with PI appear to be robust against violations of homoskedasticity.

Given its flexibility and good performance shown in multiple simulation studies (see Bauer et al., 2020; Kolbe et al., 2021), MNLFA seems to be a promising method for assessing measurement invariance with respect to categorical and continuous background variables. Performing MNLFA for measurement invariance assessment may, however, not be straightforward for researchers without access to M*plus* or SAS. In **Chapter 5**, we aim to make MNLFA more accessible by providing a detailed guideline on performing this method in the open-source R (R Core Team, 2021) package `OpenMx` (Boker et al., 2011). The chapter includes a demonstration of how MNLFA can be applied in R for evaluating measurement invariance with respect to a dichotomous and continuous background variable simultaneously. In addition to this demonstration, we show that the parameter estimates are identical to those obtained when using MNLFA in M*plus*. This provides a valuable cross-validation that both optimizers converge on the same parameter estimates.

Summary in Dutch/Samenvatting

Nieuwe manieren om meetinvariantie te onderzoeken

De meting van latente constructen zoals rekenvaardigheid en sociale angst speelt een belangrijke rol binnen de sociale en gedragswetenschappen. Omdat dergelijke constructen niet direct kunnen worden gemeten, wordt er vaak gekeken naar observeerbare indicatoren. Voor een zinvolle vergelijking tussen individuen of groepen, moet de relatie tussen elke geobserveerde indicator en het latente construct hetzelfde zijn voor alle individuen of groepen. Deze assumptie wordt ook wel meetinvariantie genoemd. Als er sprake is van meetinvariantie met betrekking tot een specifieke achtergrondvariabele, heeft die achtergrondvariabele geen invloed op de meting van het latente construct. Als de assumptie van meetinvariantie geschonden wordt, kunnen verschillen in geobserveerde scores tussen individuen of groepen het gevolg zijn van verschillen in de achtergrondvariabele in plaats van verschillen in het latente construct wat gepoogd wordt te meten. Het is daarom belangrijk om meetinvariantie te toetsen voordat individuen of groepen op latente constructen worden vergeleken.

Een veel gebruikte klasse van methoden om meetinvariantie te toetsen binnen structurele vergelijkingsmodellen is *confirmatory factor analysis* (CFA). Een van de traditionele CFA methoden om meetinvariantie met betrekking tot een categorische achtergrondvariabele te onderzoeken is *multiple-group* CFA (MGCFA; Vandenberg & Lance, 2000). Bij MGCFA wordt een CFA model voor elke groep geschat en wordt meetinvariantie getoetst door modellen met en zonder gelijkheidsrestricties op de parameters te vergelijken. Naast MGCFA zijn er ook *single-group* methoden voor het onderzoeken van meetinvariantie, waaronder *restricted factor analysis* (RFA; Oort, 1992) en *moderated nonlinear factor analysis* (MNLFA; Bauer & Hussong, 2009). Bij *single-group* methoden wordt er slechts één CFA model geschat voor de gehele steekproef. Deze methoden zijn dan ook geschikt voor kleinere steekproeven en voor het toetsen van meetinvariantie met betrekking tot meerdere continue en categorische achtergrondvariabelen tegelijk.

Dit proefschrift richt zich op nieuwe manieren om meetinvariantie te toetsen met behulp van *single-group* methoden. Het proefschrift begint met een onderzoek over RFA, een methode die geschikt is om uniforme meetinvariantie te toetsen, maar een aanvullende statistische techniek vereist om niet-uniforme meetinvariantie te toetsen. De aanvullende statisische techniek is nodig om de interactie tussen het latente construct en de achtergrondvariabele te modelleren. RFA wordt vaak toegepast in combinatie met *latent moderated structural equations* (LMS). Hoewel gebleken is dat LMS hoge statistische power heeft om schendingen van niet-uniforme meetinvariantie te detecteren, zijn er ook zeer hoge percentages Type I fouten waargenomen bij het gebruik van deze methode in RFA modellen (Barendse et al., 2010, 2012; Woods & Grimm, 2011). In **Hoofdstuk 2** stellen wij daarom voor om *product indicators* (PI) te gebruiken als alternatief voor LMS in RFA modellen. Aan de hand van een enkele gesimuleerde dataset laten we zien hoe de PI methode in RFA modellen kan worden gebruikt om niet-uniforme meetinvariantie te toetsen. De verkregen resultaten zijn vergelijkbaar met die van LMS,

172

wat aangeeft dat de PI methode een geschikt alternatief kan zijn.

Het gebruik van PI in RFA modellen wordt uitvoeriger onderzocht met een simulatiestudie in **Hoofdstuk 3**. In dit hoofdstuk evalueren we de Type I fouten en de statistische power van de LMS en PI methoden om schendingen van uniforme en niet-uniforme meetinvariantie te detecteren. De resultaten van de simulatiestudie laten zien dat de PI methode in bijna alle simulatiecondities een vergelijkbare statistische power maar lagere percentages Type I fouten heeft dan LMS. De percentages Type I fouten van PI liggen in alle condities dicht bij het nominale significantieniveau. Dit suggereert dat het gebruik van PI in RFA modellen de kans op foute conclusies met betrekking tot welke geobserveerde indicatoren de assumptie van uniforme of niet-uniforme meetinvariantie schenden kan minimaliseren. In overeenstemming met eerdere studies (Barendse et al., 2010, 2012; Woods & Grimm, 2011), worden zeer hoge percentages Type I fouten waargenomen bij het gebruik van LMS in RFA modellen.

Een mogelijke verklaring voor de te hoge Type I foutenpercentages van LMS, is een schending van de assumptie van homoskedasticiteit (anders gezegd, gelijke factor en residuele varianties; Chun et al., 2016; Meredith & Teresi, 2006). Om deze mogelijke verklaring te bevestigen of te ontkrachten, onderzoeken wij in **Hoofdstuk 4** het effect van schendingen van deze assumptie op de prestaties van RFA in combinatie met LMS en PI. In tegenstelling tot RFA is het voor de recent voorgestelde MNLFA methode (Bauer & Hussong, 2009; Bauer, 2017) voor het toetsen van meetinvariantie niet nodig om homoskedasticiteit ten opzichte van de achtergrondvariabele te veronderstellen. In dit hoofdstuk wordt er daarom ook een vergelijking tussen RFA en MNLFA gemaakt onder elk van de verschillende simulatiecondities. Uit de resultaten van de simulatiestudie blijkt dat de percentages Type I fouten van RFA in combinatie met LMS aanzienlijk toenemen als functie van heteroskedasticiteit, terwijl MNLFA en RFA in combinatie met PI robuust blijken te zijn tegen schendingen van homoskedasticiteit.

Gezien de flexibiliteit en de goede prestaties die zijn aangetoond in eerdere onderzoeken (Bauer et al., 2020; Kolbe et al., 2021), lijkt MNLFA een veelbelovende methode te zijn voor het toetsen van meetinvariantie met betrekking tot categorische en continue achtergrondvariabelen. Het uitvoeren van MNLFA is echter niet eenvoudig voor onderzoekers zonder toegang tot M*plus* of SAS. In **Hoofdstuk 5** proberen we MNLFA toegankelijker te maken door een gedetailleerd stappenplan te presenteren voor het uitvoeren van deze methode in het open-source R (R Core Team, 2021) pakket `OpenMx` (Boker et al., 2011). Het hoofdstuk bevat een illustratie van hoe MNLFA kan worden toegepast in R om meetinvariantie ten opzichte van een dichotome en continue achtergrondvariabele tegelijk te toetsen. Naast deze illustratie vergelijken we de resultaten verkregen in R met de resultaten verkregen in M*plus* en vinden we identieke parameterschattingen. Dit levert een waardevolle kruisvalidatie op, waarbij geconcludeerd kan worden dat beide manieren van het uitvoeren van MNLFA convergeren naar dezelfde parameterschattingen.

Acknowledgements/Dankwoord

Dit proefschrift had ik nooit kunnen schrijven zonder de hulp en steun van begeleiders, collega's, vrienden en familie. Daarom wil ik graag afsluiten met een dankwoord. Om te beginnen bij **Frans Oort**, mijn promotor. Jij stelde je als begeleider toegankelijk en enthousiast op, waardoor ik altijd bij jou durfde aan te kloppen met al mijn inhoudelijke en persoonlijke vragen – en dat waren er behoorlijk wat. Jij gaf mij telkens weer het vertrouwen dat het wel goed zou komen. Dankjewel voor alles wat ik van jou heb mogen leren! Dan **Terrence Jorgensen**, één van mijn copromotoren. Terrence, you have helped me in so many different ways. You taught me how to perform simulation studies, how to write scientific papers, how to not be discouraged by errors in R, how to celebrate achievements, but above all how not to blindly trust your beer recommendations when in Zurich. Saying that this dissertation would not be finished without your help is an understatement. Thank you so much! **Suzanne Jak**, dankjewel dat je mij als copromotor hebt aangemoedigd in al mijn keuzes. Door jou heb ik geleerd niet eindeloos na te denken en gewoon eens dingen te proberen. Je hebt dit proces daardoor niet alleen productiever, maar ook veel leuker gemaakt. Tegelijkertijd gaf je mij de ruimte om op de rem te trappen op de momenten dat dat nodig was. Ik ben je daar nog altijd dankbaar voor.

Tijdens dit promotietraject heb ik ook veel mogen leren van alle **collega's van de Methods & Statistics afdeling**. Ik wil jullie daar allemaal voor bedanken, maar in het bijzonder **Niels Smits** en **Andries van der Ark**. Hoewel ik promoveer bij Frans, Terrence en Suzanne, maken jullie wat mij betreft ook onderdeel uit van mijn 'wetenschappelijke voorouders'. Jullie hebben als begeleiders van mijn onderzoeksstage en later als collega's veel bijgedragen aan mijn ontwikkeling. Ik heb met plezier met jullie samengewerkt. **Kees Jan Kan**, dankjewel voor de gezelligheid, het luisterend oor en de nodige hoeveelheid drop op kamer D7.24. Wie ik ook graag wil bedanken is **Dylan Molenaar**. Ik weet nog goed dat ik jou ooit een mail stuurde met de vraag of je een keer kon meedenken over een onderzoeksidee van Terrence en mij. Die mail heeft uiteindelijk geleid tot een inspirerende samenwerking. Jij hebt als co-auteur mij in de juiste wetenschappelijke richting geduwd. Mijn dank gaat ook uit naar mijn **nieuwe collega's** bij Cito voor het warme welkom en de ondersteuning tijdens het afronden van dit proefschrift. Daarnaast wil ik de leden van de promotiecommissie **prof. dr. Marieke Timmerman, prof. dr. Denny Borsboom, dr. Kees Jan Kan, dr. Malthilde Verdam en prof. dr. Andries van der Ark** hartelijk danken voor het beoordelen van mijn proefschrift en het opponeren tijdens de promotieplechtigheid.

Dankzij mede-promovendi **Debby, Hannelies en Letty** voelden de werkdagen op de Universiteit van Amsterdam altijd als een warm bad. Wat begon als collega's is uitgegroeid tot een hechte vriendschap die voor mij tot op de dag van vandaag heel waardevol is. Met jullie kon ik lachen, lunchen (is het al 12 uur?), hard werken en ontspannen, maar ook poggen (proefschriftontwijkend gedragen), ventileren, klagen en huilen. **Debby**, dankjewel voor alle kledingmerktips, voor het kunnen delen van mijn liefde voor katten en Zwolle, maar vooral voor alle mooie gesprekken die we hebben gehad samen. **Hannelies**, wat ben ik blij met zo'n lieve vriendin aan mijn zijde. Alleen al het grote aantal uren

dat wij aan de telefoon hebben gehangen getuigt van jouw gezelligheid en onuitputtelijke support. Rocyclen, borrelen, Lowlands, winkelen: op naar nog veel meer leuke herinneringen samen! **Letty**, dankjewel voor jouw enthousiasme, positiviteit en eerlijkheid. Met jou is dit avontuur ooit begonnen. Nog voordat mijn PhD startte gingen we samen naar een conferentie in Wenen. Na zo veel gezellige momenten, zoals rondstruinen door het natuurhistorisch museum en avondeten bij schnitzelparadijs Centimeter, wist ik één ding zeker: als dit mijn collega wordt mag ik van geluk spreken! Het schrijven van dit proefschrift is onlosmakelijk verbonden met jou en had ik met niemand anders willen meemaken.

Omdat het voor mij zo vanzelfsprekend voelt dat mijn vrienden een onderdeel van mijn leven zijn, vergeet ik soms bijna hoeveel ze voor mij betekenen. Dank dat jullie voor de gezonde afleiding hebben gezorgd. Van samen proosten op het nieuwe jaar tot Koningsdag vieren in ons oude appartement in Amsterdam. Ik heb hier dankzij jullie zo veel fijne herinneringen aan. Een paar vriendinnen wil ik in het bijzonder bedanken. **Julia van Leeuwen**, we zagen elkaar voor het eerst tijdens de introductiedag van onze masteropleiding in Amsterdam. Omdat jij nog in Groningen woonde logeerde je de eerste weken bij mij en daardoor ontstond er in korte tijd een hele hechte band. Even later werd onze vriendschap uitgebreid met lieve **Annemieke en Tessa**. Dagjes weg, musicals, escape rooms, uitgebreid dineren, een bruiloft en de komst van de eerste UvA Girlz baby: ik hoop nog veel meer mooie dingen met jullie mee te mogen maken. **Sophie, Frederike, Laura, José, Mirte, Veerle, Milou, Jonne, Sanne en Loes**, er gaat niks boven Groningen maar vooral niks boven mijn liefde voor jullie. Elk moment met jullie leidt tot een mooie herinnering. Bedankt voor alles! Lieve **Julia van Ittersum**, ik ben heel dankbaar dat we al meer dan 15 jaar zo'n onvoorwaardelijke vriendschap hebben. Jij kent mij als geen ander en hebt aan een half woord al genoeg om te begrijpen wat ik bedoel. Dankjewel dat ik altijd bij je terecht kan. Ik hoop dat we onze stedentriptraditie snel weer kunnen voortzetten.

Tot slot mijn allerliefste familie. Zij hielpen mij gedurdende dit traject te beseffen dat er veel belangrijkere dingen in het leven zijn dan dit proefschrift. **Karel, Els, Rein, Eva en Thom**, ik ben dankbaar dat ik heb mogen opgroeien met zulke lieve opa's en oma's om mij heen. **Patrick en Eline**, dank jullie wel voor jullie steun en de vrolijke invloed van jullie kinderen. Op naar nog veel meer gezelligheid samen! **Paul, Marianne, Ruben en Maartje**, ik had nooit durven hopen dat een schoonfamilie zo als thuis zou voelen. In onze jaren samen is er van alles gebeurd en veranderd, maar dankzij jullie bleef er ook veel hetzelfde. Dat ik altijd kan rekenen op jullie dosis liefde, humor en betrokkenheid is een hele geruststellende gedachte. **Thomas**, als broer en zus hoeven we elkaar niet veel te spreken om te weten dat we altijd bij elkaar kunnen aankloppen. Jij hebt mij tijdens het schrijven van dit proefschrift geleerd om auto te rijden. Het bleek bijna lastiger dan een PhD afronden, maar het is gelukt. Dan mijn ouders **Jeroen en Diana**, zonder jullie had ik dit nooit gekund. Jullie hebben mij de vrijheid gegeven om te doen wat ik wil, mij laten vertrouwen op mijn gevoel en mij geleerd om te genieten van de natuur en muziek.

Zoals in één van de vele liedjes die we elkaar altijd aanraden wordt gezongen: 'Reis ver, drink wijn, denk na, lach hard, duik diep, kom terug'. Waar ik ook heen ga, jullie zijn mijn thuis.

Lieve **Julian**, de laatste woorden van mijn proefschrift zijn voor jou. Ik kan me voorstellen dat jij minstens net zo opgelucht bent als ik dat dit avontuur er nu op zit. Ik heb zo vaak mijn presentaties, artikelen en samenvattingen aan jou mogen voorleggen dat je inmiddels bovengemiddeld veel weet van meetinvariantie en hoe je *heteroskedasticity* uitspreekt. Maar ook op niet-inhoudelijk vlak was niets voor jou te veel: wasjes draaien, peptalks geven, koken, relativeren, mijn favoriete series meekijken, geklaag verdragen, oplossingen zoeken, samen hardlopen, een luisterend oor bieden, chocolade muffins kopen, afleiden, ontstressen, en nog zo veel meer. Ik ben je daar onmetelijk dankbaar voor. Laatst was het 10 jaar geleden dat we elkaar voor het eerst hebben ontmoet en ik kan me die ontmoeting nog goed herinneren. Waar ik toen gelijk onder de indruk was van de dronken pianoserenade, ben ik dat nu vooral van jouw geduld, positiviteit en liefde voor iedereen om je heen. Nu dit proefschrift af is, is het eindelijk weer tijd om samen te genieten van een Vietnamese ijskoffie aan het strand!