University of Amsterdam

UvA-DARE (Digital Academic Repository)

The Impact of Unmodeled Heteroskedasticity on Assessing Measurement Invariance in Single-group Models

Kolbe, L.; Jorgensen, T.D.; Molenaar, D.

DOI
10.1080/10705511.2020.1766357

Publication date
2021

Document Version
Final published version

Published in
Structural Equation Modeling

License
CC BY-NC

Link to publication

Citation for published version (APA):
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. https://doi.org/10.1080/10705511.2020.1766357

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:10 Mar 2023

STRUCTURAL EQUATION MODELING: A MULTIDISCIPLINARY JOURNAL
2021, VOL. 28, NO. 1, 82–98
https://doi.org/10.1080/10705511.2020.1766357

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

The Impact of Unmodeled Heteroskedasticity on Assessing Measurement Invariance in Single-group Models

Laura Kolbe, Terrence D. Jorgensen, and Dylan Molenaar

University of Amsterdam

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.

KEYWORDS

Heteroskedasticity; measurement Invariance; moderated Nonlinear Factor Analysis; restricted Factor Analysis; product Indicators

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

$$f_1(X|T,V) = f_2(X|T), \tag{1}$$

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 paper) measuring the latent construct of interest T, and V is a set of background variables that is 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; Finch & French, 2018; Horn & McArdle, 1992; Marsh, 1994), as described below. But much less advice is available for 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 (Marsh et al., 2017; B. Muthén & Asparouhov, 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 categorical 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

CONTACT Laura Kolbe l.kolbe@uva.nl University of Amsterdam.

[1] Exceptions include Kolbe and Jorgensen (2019), Masyn (2017), and Suh (2015).

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

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 imply 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 a 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 and 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 categorized, 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 and Hussong (2009, but see the earlier work by, e.g., Neale, 1998; Neale, Aggen, Maes, Kubarych and Schmitt, 2006; Mehta and 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 model 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 and 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; Harpole, 2015; Kolbe & Jorgensen, 2019), 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 and Jorgensen (2019) did not observe inflated Type I error rates despite common-factor variances being unequal. The Appendix 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 paper 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 paper concludes with advice for applied researchers and suggestions for future research.

Single-Group Models

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 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

$$x_i = \tau + \Lambda t_i + \varepsilon_i, \tag{2}$$

where x_i is a $J \times 1$ vector of J observed indicator scores for person i, τ is a $J \times 1$ vector of indicator intercepts, Λ is a $J \times 1$ vector of factor loadings, t_i is the common-factor score for person i and ε_i is a $J \times 1$ 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 Λ with respect to V, and scalar invariance additionally requires equal intercepts τ. In order to evaluate metric and scalar measurement invariance in a single-group model, the model for continuous indicators X can be rewritten as

$$\begin{aligned} x_i &= \tau_i + \Lambda_i t_i + \varepsilon_i \\ &= (\tau_0 + b v_i) + (\Lambda_0 + c v_i) t_i + \varepsilon_i, \end{aligned} \tag{3}$$

where v_i is the background variable score for person i, τ_0 is a $J \times 1$ vector of baseline intercepts when person i's score on the variable V is $v_i = 0$, and Λ_0 is a $J \times 1$ vector of baseline factor loading when $v_i = 0$. The $J \times 1$ vectors b and c are of special interest because they contain coefficients that reflect violations of measurement invariance (i.e., DIF). A nonzero element in 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 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 b and 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 b and 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 b and 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 appproaches share the same general form (Equation (3)), but differ in the way the background variable V is modeled and b and 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.

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 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 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 c – the $T \times V$ interaction effect 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 and 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 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 1 shows an RFA model amenable to LMS for assessing measurement invariance with respect to variable V. In this example, T is measured by J 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 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 1 because they are not estimated, but the estimation implicitly allows those associations to exist. A nonzero effect of V on X_j, denoted b_j, implies uniform DIF for indicator j, whereas a nonzero effect of $T \times V$ on X_j, denoted c_j, implies nonuniform DIF for indicator j.

The LMS approach is a full information maximum likelihood approach that assumes 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 (Dimitruk et al., 2007; A. Klein & Moosbrugger, 2000), but with nonnormal variables inflated Type I error rates were observed when testing for the

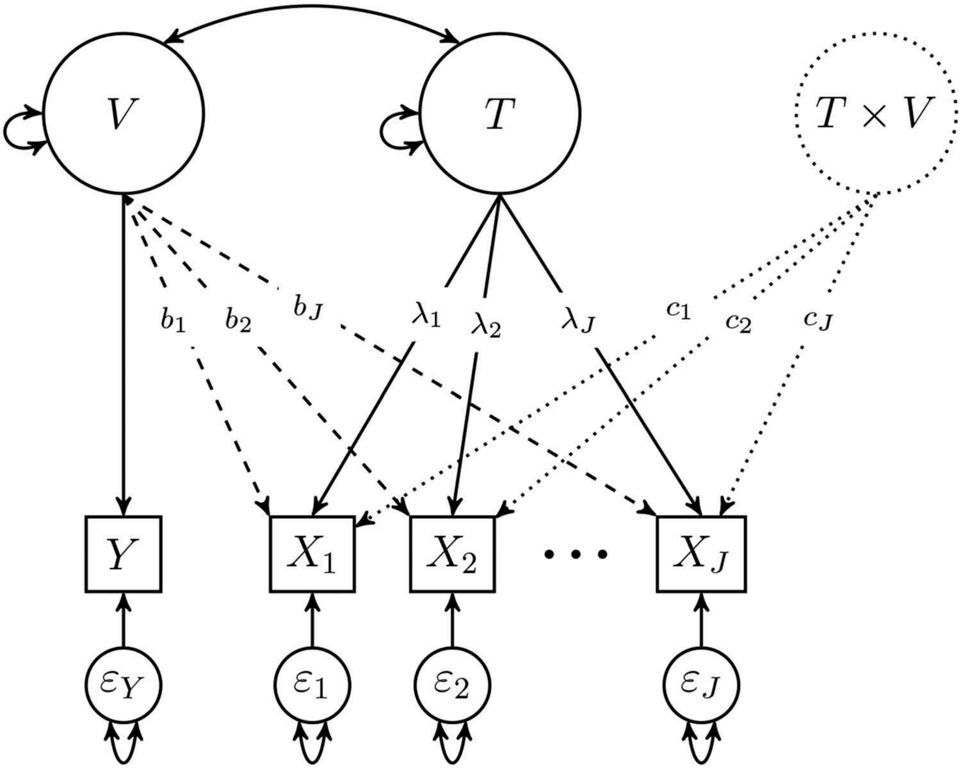

Figure 1. An RFA model with LMS for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to assess scalar and metric invariance, respectively.

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

Figure 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 2, the indicators of T and V are mean-centered. The double-mean-centered product indicator of the j-th indicator is denoted $(Y^C \times X_j^C)^C$. 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 the Appendix 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, 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 (Lin et al., 2010; Marsh et al., 2004).

Moderated Nonlinear Factor Analysis

The MNLFA approach (Bauer, 2017; Bauer & Hussong, 2009) includes the background variable V in the model only as a moderator variable, whereby parameters can be defined as functions of V. Figure 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 (3)) as

$$\tau_i = \tau_0 + b v_i, \qquad (4)$$

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

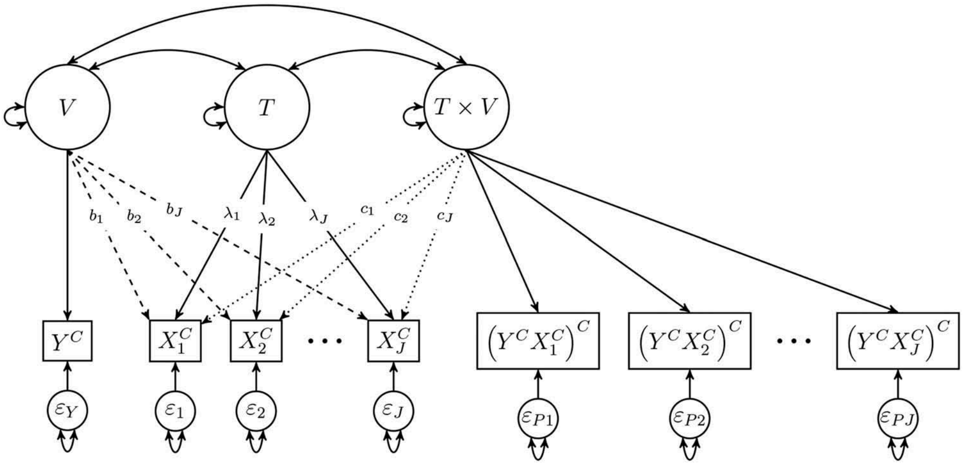

Figure 2. An RFA model with PI for assessing measurement invariance. The dashed and dotted arrows represent effects that may be estimated to assess scalar and metric invariance, respectively.

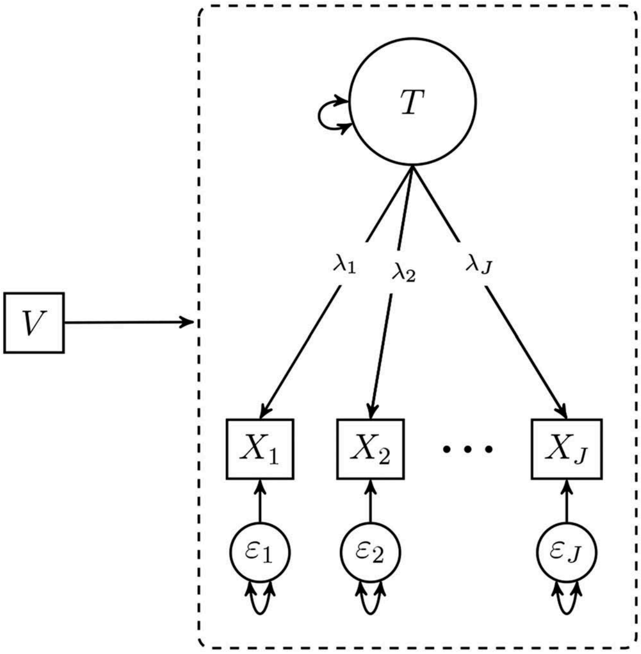

Figure 3. An MNLFA model for assessing measurement invariance. The variable V may have an effect on all parameters in the model represented in the dashed border.

$$\mathbf{\Lambda}_i = \mathbf{\Lambda}_0 + c v_i, \tag{5}$$

where any nonzero element of c reflects a linear change in $\mathbf{\Lambda}$ 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

$$\alpha_i = \alpha_0 + \Gamma v_i. \tag{6}$$

Here α_0 is the baseline common-factor mean when $v_i = 0$ and Γ 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 and Hussong (2009) proposed to model variances as exponential functions of V. The variance of the common factor T may be written as

$$\psi_i = \psi_0 \exp(\beta v_i), \tag{7}$$

where ψ_0 is the baseline common-factor variance when $v_i = 0$ and β 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,

$$\varepsilon_i = \varepsilon_0 \exp(\delta v_i), \tag{8}$$

where ε_0 is a vector of baseline residual variances and the effects of V on the residual variances are captured by δ. The baseline coefficients for the common factor α_0 and ψ_0 can be fixed at 0 and 1, respectively, in order to identify the model in the situation that an anchor indicator's intercept and loading are not constrained to 0 and 1 for identification.

Although MNLFA and RFA differ in the way V is modeled and b and c are estimated, they share the same general model for the indicators X (Equation (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 Mehta and Neale (2005), Molenaar et al. (2010), Neale (1998), Neale et al. (2006), and Purcell (2002).

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).

Data Generation

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

$$x_i = \tau + \mathbf{\Lambda} t_i + b v_i + c t_i v_i + \varepsilon_i \tag{9}$$

where 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 ε_i is a vector of 10 residual scores of person i. Moreover, the vector τ includes 10 intercepts set at 0 for all indicators, $\mathbf{\Lambda}$ includes 10 common factor loadings set at 0.8 for all indicators, and b and 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 V. 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, $\exp(-0.25v_i)$, or $\exp(0.25v_i)$. Hence, there were three levels of common-factor heteroskedasticity: $\beta = -0.25$, $\beta = 0$ (i.e., homoskedasticity), and $\beta = 0.25$. Figure 4 shows the common-factor variances as a function of V for different levels of β. 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 $-3 \leq V \leq 3$.

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: $\delta = -0.25$, $\delta = 0$ (i.e.,

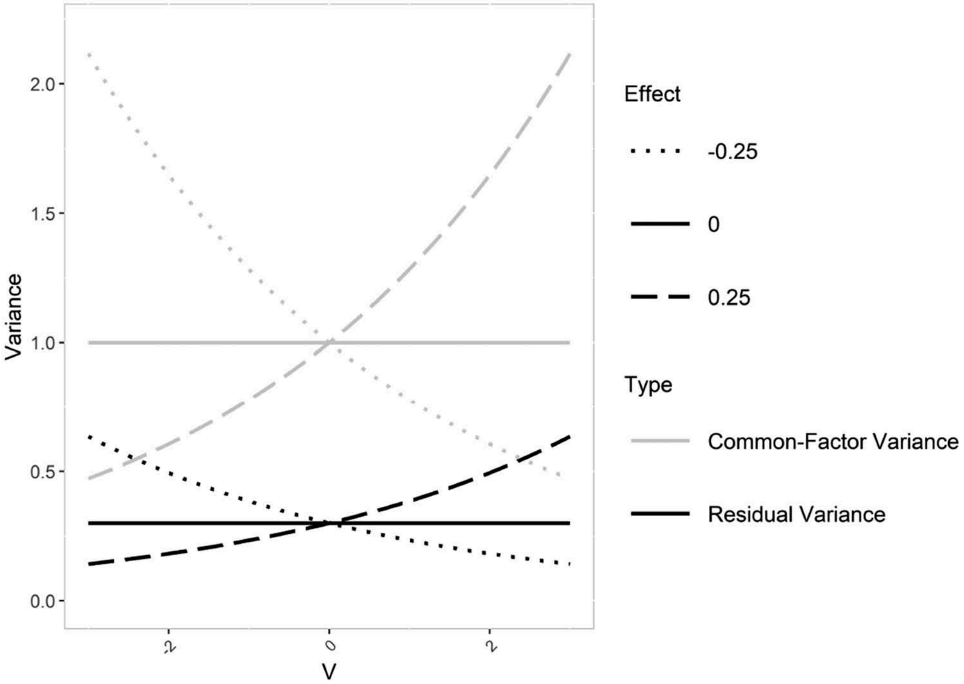

Figure 4. The common-factor and residual variances as a function of continuous V.

homoskedasticity), and $\delta = 0.25$. Figure 4 shows the residual variances as a function of V for different levels of δ. 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 the 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 the online supplementary materials.

Categorical V. In conditions where the background variable V is a categorical variable, we generated a dummy code that represents 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: $\beta = \ln(0.5)$, $\beta = 0$ (i.e., homoskedasticity), $\beta = \ln(1.5)$, and $\beta = \ln(2)$.

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: $\delta = \ln(0.15/0.3)$, $\delta = 0$ (i.e., homoskedasticity), and $\delta = \ln(0.6/0.3)$.

A violation of scalar invariance of the second and third indicators was introduced by fixing b at 0.5, and a violation of metric invariance of the fourth and fifth indicators 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).

Analysis

When measurement invariance was examined with RFA, an unconstrained model was fitted in which all elements in b and c were freely estimated, except for the elements corresponding to the ninth and tenth indicators. 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 j's $b_j = 0$ and $c_j = 0$ using a 2-*df* Wald test with $\alpha = .05$ level of significance. In order to enable the estimation of the b and 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-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 indicators 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 0 and 1, 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 b and c of the indicators with DIF were evaluated for each method by calculating the relative bias, root mean squared error

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

[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. But we did conduct a set of example analyses applied to real data, available on our Open Science Framework project https://osf.io/vsp4f/, which showed the results from a homoskedastic MNLFA and RFA/LMS yielded very similar results.

(RMSE), and coverage rates. The relative bias of the parameter estimate b of Indicator 2 was defined as a percentage using $((\bar{b} - b)/b) * 100\%$, 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 the online supplementary materials. 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 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.

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 the online supplementary materials. 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 and 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.

Continuous V

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 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 the online supplementary materials. 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 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 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 $\beta = 0$ (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 $\beta = 0$.

In conditions with common-factor heteroskedasticity (i.e., $\beta = -0.25$ 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., $\beta = 0.25$ and $\delta = 0.25$), large inflation of the error rates of RFA/LMS was observed, and the inflation was exacerbated in larger samples. In contrast, when the effects on the variances were in opposite directions (e.g., $\beta = -0.25$ and $\delta = 0.25$), 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.

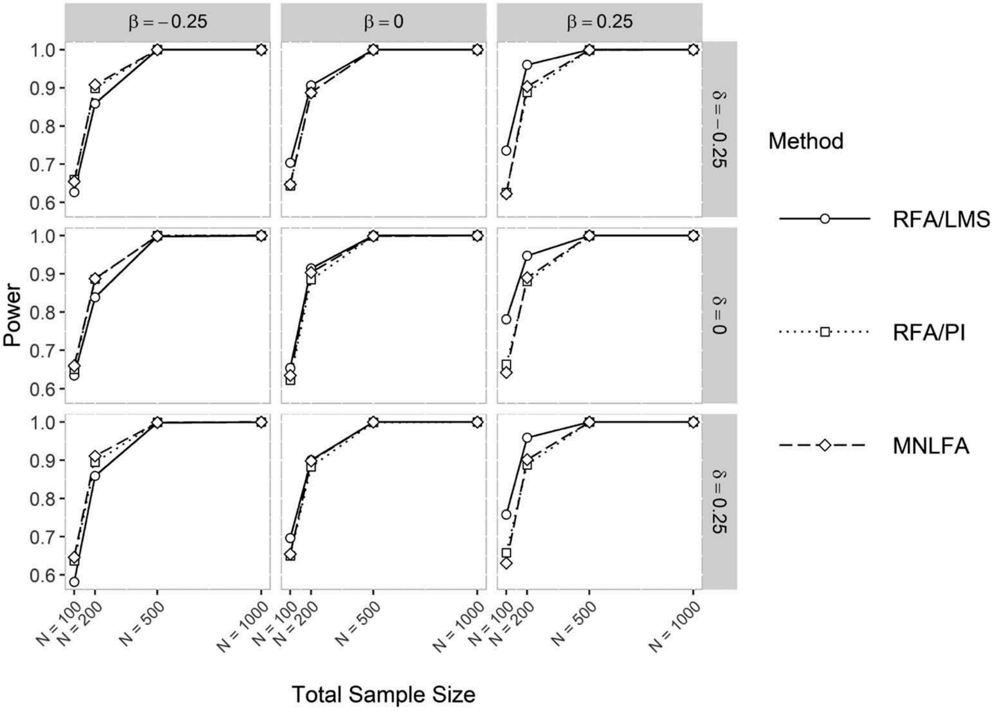

Figure 5. The power to detect a violation of metric invariance of Indicator 4 (i.e., $c_4 \neq 0$) of each method across all conditions with a continuous V. Note that β is the effect of V on the common-factor variance, and δ is the effect of V on the indicator's residual variance.

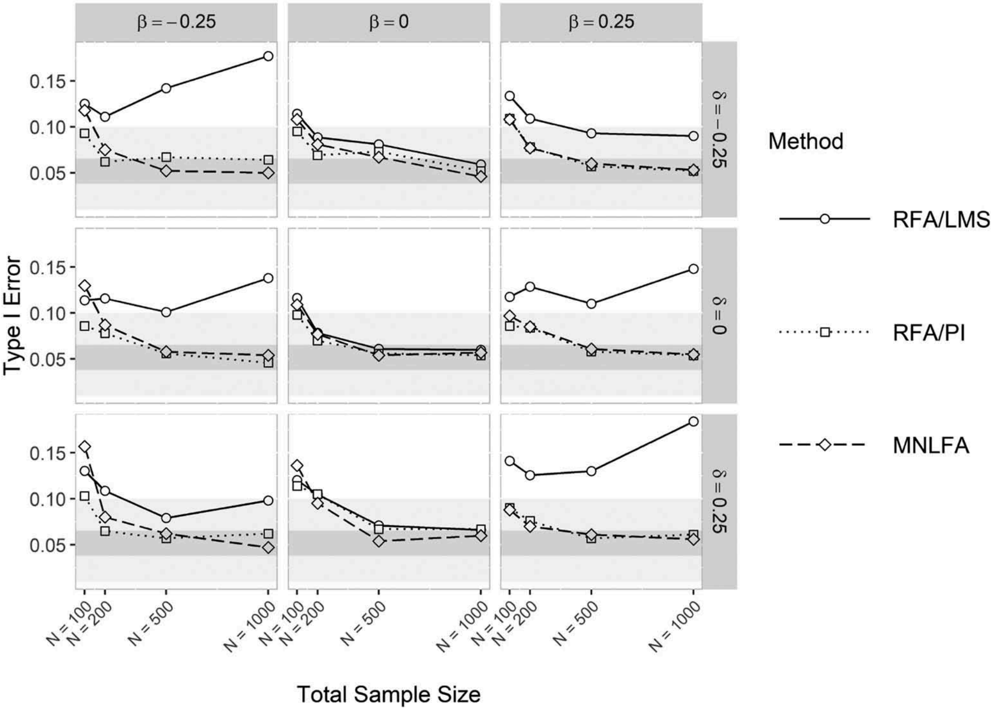

Figure 6. The Type I error rates for Indicator 1 of each method across all conditions with a continuous V. Note that β is the effect of V on the common-factor variance, and δ is the effect of V on the indicator's residual variance.

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 the online supplementary materials. 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 the largest when β 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 the online supplementary materials. 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 $\beta = -0.25$ 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 the online supplementary materials. Overall, all methods showed acceptable coverage rates (always >0.90) for the parameter estimate b. The RFA/PI method obtained coverage rates closest to 0.95 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 0.90 for c, RFA/PI and RFA/LMS frequently obtained unacceptable coverage rates. For both methods, the coverage rates for c decreased with N and β. The RFA/LMS method obtained the lowest coverage rates for the parameter estimate c. The lowest coverage rate of 0.07 was obtained when $\beta = 0.25$, $\delta = -0.25$, and $N = 1000$.

Categorical V

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 the online supplementary materials. 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 0.83 to 0.98, where a negative effect on the residual variance led to a 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 7 shows the power of methods to detect violations of metric invariance. In conditions with equal common-factor variances across groups (i.e., $\beta = 0$), 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., $\beta = \ln(1.5)$ or $\beta = \ln(2)$), but performed substantially worse when the focal group's common-factor variance was smaller (i.e., $\beta = \ln(0.5)$).

The Type I error rates across all conditions with a categorical V are illustrated in Figure 8. Note that we specified y-axis limits of 0 and 0.15 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 errors, 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 the online supplementary materials for exact error rates). This inflation was smallest 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 $\beta = \ln(0.5)$, $\delta = \ln(0.15/0.3)$, 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 β and δ 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 the online supplementary materials. 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

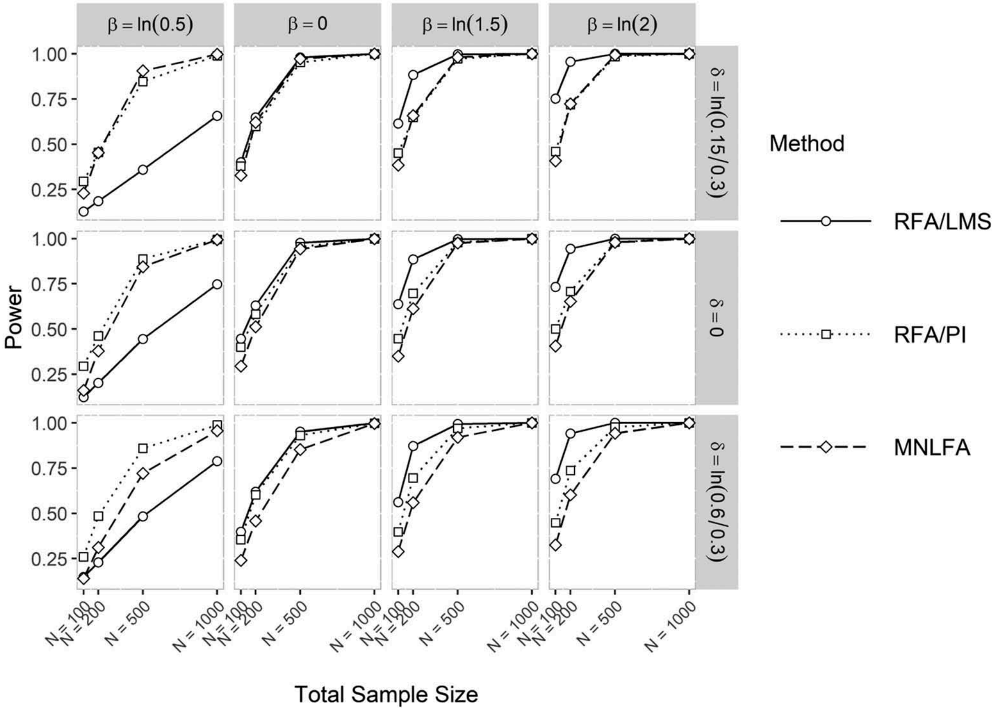

Figure 7. The power to detect a violation of metric invariance of Indicator 4 (i.e., $c_4 \neq 0$) of each method across all conditions with a categorical V. Note that β is the effect of V on the common-factor variance, and δ is the effect of V on the indicator's residual variance.

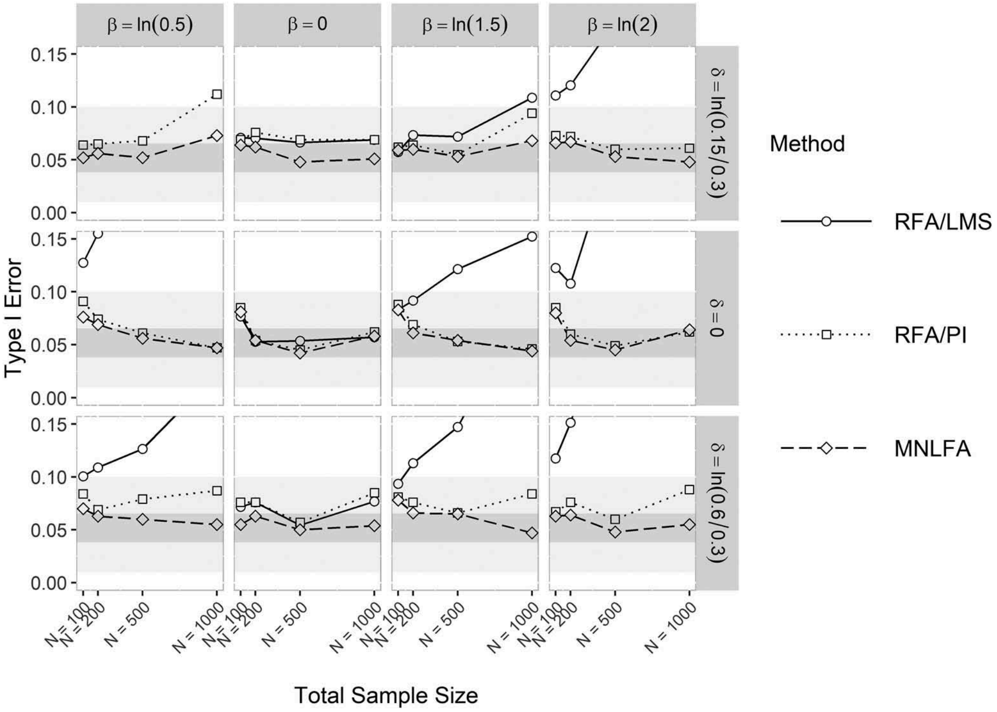

Figure 8. The Type I error rates for Indicator 1 of each method across all conditions with a categorical V. Note that β is the effect of V on the common-factor variance, and δ is the effect of V on the indicator's residual variance. The y axis stops at 0.15 in order to allow a detailed comparison of methods with (nearly) nominal error rates, but note that it prevents plotting some extremely inflated error rates in certain conditions of RFA/LMS.

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 β and δ 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 the online supplementary materials. 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., $\beta \neq 0$ or $\delta \neq 0$). The RFA/PI method generally obtained the second lowest RMSE of the parameter estimate c when β was positive, but obtained the highest RMSE when β was negative and δ was positive. In the conditions with common-factor homoskedasticity (i.e., $\beta = 0$), 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 the online supplementary materials. Similar to conditions with a continuous V, all methods obtained acceptable coverage rates for the parameter estimate b (always >0.90), 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 the parameter estimate c. For RFA/PI, the coverage rates for c where acceptable (i.e., larger than 0.80) in smaller sample-size conditions, but were frequently unacceptable (i.e., smaller than 0.80) 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 0.01 at its lowest.

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 the online supplementary materials. 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.

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 a high power to detect violations of measurement invariance, but severely inflated Type I error rates have also been observed (Barendse et al., 2010, 2012; Kolbe & Jorgensen, 2019; Woods & Grimm, 2011). 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 were 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 the Appendix, 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 SAS, 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 SAS and 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 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 (Strobl et al., 2015; Tutz & Berger, 2016), 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, 2015; Molenaar et al., 2012, 2011; Molenaar et al., 2010), and stochastic process-based testing (Merkle et al., 2014; Merkle & Zeileis, 2013). 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.

ORCID

Laura Kolbe http://orcid.org/0000-0002-4285-3939
Terrence D. Jorgensen http://orcid.org/0000-0001-5111-6773
Dylan Molenaar http://orcid.org/0000-0002-7168-3238

References

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

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 with latent moderated structures to detect uniform and nonuniform measurement. *Advances in Statistical Analysis, 94,* 117–127. https://doi.org/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*, 561–579. https://doi.org/10.1080/10705511.2012.713261

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*, 507–526. https://doi.org/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*, 43–55. https://doi.org/10.1080/10705511.2019.1642754

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

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., & Brick, T.; others. (2011). *OpenMx: An open source extended structural equation modeling framework.* An open source extended structural equation modeling framework.

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

Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: An expository note. *The American Statistician*, *36*, 153–157. https://doi.org/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*, 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for Chun, S., Stark, S., Kim, E. S., & Chernyshenko, O. S. (2016). MIMIC methods for procedure. *Applied Psychological Measurement*, *40*, 486–499. https://doi.org/10.1177/0146621616659738

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

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

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

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

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

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling*, *25*, 621–638. https://doi.org/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*, 82–109. https://doi.org/10.1080/10705510903439003

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*, 257–258. https://doi.org/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*, 117–144. https://doi.org/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. comparison of latent variables: Equivalence testing and a projection-based approach.

Jöreskog, K. G., & Yang,. (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). Lawrence Erlbaum Associates.

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

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*, 201. https://doi.org/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*, 457–474. https://doi.org/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*, 647–673. https://doi.org/10.1080/00273170701710205

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. https://doi.org/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. https://doi.org/10.3758/s13428-018-1151-3

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*, 374–391. https://doi.org/10.1080/10705511.2010.488999

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*, 497–519. https://doi.org/10.1207/s15328007sem1304_1

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*, 637–665. https://doi.org/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*, 19–40. https://doi.org/10.1037/1082-989X.7.1.19

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

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2017). 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*, 524–545. https://doi.org/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*, 275–300. https://doi.org/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, 180–197. https://doi.org/10.1080/10705511.2016.1254049

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

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

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

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*, S69–S77. https://doi.org/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*, 569–584. https://doi.org/10.1007/S11336-013-9376-7

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*, 59–82. https://doi.org/10.1007/S11336-012-9302-4

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

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*, 455–478. https://doi.org/10.1007/S11336-012-9273-5

Molenaar, D., Dolan, C. V., & van der Maas, H. L. (2011). Modeling ability differentiation in the second-order factor model. *Structural Equation Modeling*, *18*, 578–594. https://doi.org/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*, 293–317. https://doi.org/10.1348/000711009X456935

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

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

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide (7th ed.) [Computer software manual]*. 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). Erlbaum.

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*, 1010–1034. https://doi.org/10.1016/j.addbeh.2006.03.047

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

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

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

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

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

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

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316. https://doi.org/10.1007/S11336-013-9388-3

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, 568–580 https://doi.org/10.1080/10705511.2014.937669

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

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*, 585–598. https://doi.org/10.1080/10705511.2016.1266267

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

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*, 4–70. https://doi.org/10.1177/109442810031002

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

Appendix

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

$$x_{ij} = \tau_j + \lambda_j t_i + \varepsilon_{ij}, \tag{A1}$$

where x_{ij} is the observed indicator score of person $i = 1, \ldots, N$ on indicator $j = 1, \ldots, J$, τ_j is an intercept, λ_j is a factor loading, t_i is a common factor score and ε_{ij} 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:

$$T = \sigma T', \tag{A2}$$

where

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

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

$$x_{ij} = \tau_j + \lambda_j \sigma t'_i + \varepsilon_{ij}, \tag{A4}$$

where

$$\sigma = \sigma_0 + \sigma_1 V. \tag{A5}$$

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

$$x_{ij} = \tau_j + \lambda_j(\sigma_0 t'_i + \sigma_1 v_i t'_i) + \varepsilon_{ij}, \tag{A6}$$

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.