Sensitivity Analysis of Longitudinal Measurement Non-Invariance:

A Second-Order Latent Growth Model Approach

with Ordered-Categorical Indicators

by

Yu Liu

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved June 2016 by the
Graduate Supervisory Committee:

Jenn-Yun Tein, Co-Chair
Stephen G. West, Co-Chair
Samuel Green
Kevin J. Grimm

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

Researchers who conduct longitudinal studies are inherently interested in studying individual and population changes over time (e.g., mathematics achievement, subjective well-being). To answer such research questions, models of change (e.g., growth models) make the assumption of longitudinal measurement invariance. In many applied situations, key constructs are measured by a collection of ordered-categorical indicators (e.g., Likert scale items). To evaluate longitudinal measurement invariance with ordered-categorical indicators, a set of hierarchical models can be sequentially tested and compared. If the statistical tests of measurement invariance fail to be supported for one of the models, it is useful to have a method with which to gauge the practical significance of the differences in measurement model parameters over time. Drawing on studies of latent growth models and second-order latent growth models with continuous indicators (e.g., Kim & Willson, 2014a; 2014b; Leite, 2007; Wirth, 2008), this study examined the performance of a potential sensitivity analysis to gauge the practical significance of violations of longitudinal measurement invariance for ordered-categorical indicators using second-order latent growth models. The change in the estimate of the second-order growth parameters following the addition of an incorrect level of measurement invariance constraints at the first-order level was used as an effect size for measurement non-invariance. This study investigated how sensitive the proposed sensitivity analysis was to different locations of non-invariance (i.e., non-invariance in the factor loadings, the thresholds, and the unique factor variances) given a sufficient sample size. This study also examined whether the sensitivity of the proposed sensitivity analysis depended on a

number of other factors including the magnitude of non-invariance, the number of non-invariant indicators, the number of non-invariant occasions, and the number of response categories in the indicators.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Researchers who conduct longitudinal studies are inherently interested in studying individual and population changes over time (e.g., mathematics achievement, depression, externalizing behavior, subjective well-being). To answer such research questions, models of change (e.g., growth models) make the assumption of longitudinal measurement invariance, i.e., the instrument reflects the same construct measured on the same scale over all time points under study and over all individuals. It is common practice for researchers to administer the same questionnaire, survey, or scale to participants and assume that longitudinal measurement invariance holds. However, in many cases this assumption may not be appropriate because the same measurement instrument can reflect a different construct at different ages (e.g., rapid changes/transitions occurring in adolescence can lead to different interpretations of the survey questions). If longitudinal measurement invariance does not hold, then the observed changes may reflect change in the properties of the measurement instrument, rather than the latent construct that the researcher intends to study. Thus, in order to draw valid conclusions about growth and change in the latent constructs of interest over time, longitudinal measurement invariance must be evaluated.

With continuous measured indicators, procedures for evaluating longitudinal measurement invariance have been developed under a confirmatory analysis (CFA) framework (e.g., Khoo, West, Wu, & Kwok, 2006; Meredith & Horn, 2001; Widaman, Ferrer, & Conger, 2010) and since then there has been several studies demonstrating the application of these procedures to empirical data sets (e.g., Millsap & Cham, 2012).

However, in many applied situations, the measured indicators are not strictly continuous. Instead, the measured indicators are often ordered-categorical (e.g., self-report or observer-report Likert scale items). These indicators are typically viewed as an ordinal outcome of a continuous underlying propensity (Bollen & Curran, 2006, p. 230). There have been a number of simulation studies examining features that can influence the degree to which CFA models assuming continuously scaled indicators can adequately model ordered-categorical indicators and result in negligible bias in parameter estimates. These studies find that it may sometimes be acceptable to treat ordered-categorical indicators as continuous, specifically when there are five or more response categories and when each of the response categories is well populated (e.g., Beauducel & Herzberg, 2006; DiStefano, 2002; Dolan, 1994; Rhemtulla, Brosseau-Liard, & Savalei, 2012). However, when there are fewer categories or when the observed distributions of the ordered-categorical indicators are skewed, treating ordered-categorical indicators as continuous can lead to biased parameter estimates. Thus, when measured indicators are ordinal, the approach of choice is often to use models that treat them as ordinal (e.g., CFA models for ordered-categorical indicators; Muthén, 1984; Wirth & Edwards, 2007).

The test of longitudinal measurement invariance with continuous indicators involves fitting and comparing a set of hierarchical models (configural vs. weak; weak vs. strong; strong vs. strict invariance; Khoo et al., 2006; Meredith & Horn, 2001). Paralleling this, with ordered-categorical indicators, a set of four hierarchical models can be sequentially tested and compared to evaluate longitudinal measurement invariance (Liu et al., in press). The configural invariance model tests the hypothesis that the general pattern of factor loadings is the same over time. The loading invariance model adds the

constraint that factor loadings are equal over time. The threshold invariance model further adds the constraint that the threshold level of going from one response category to the next is equal over time for all indicators. The unique factor invariance model adds the further constraint that all unique factor variances (and any non-zero within-wave unique factor covariance) are equal over time. Each level of longitudinal measurement invariance is associated with specific properties, which is discussed in more detail in a later section.

Statistical tests are used to compare the hierarchical models used in the test of longitudinal measurement invariance with ordered-categorical indicators. If the configural invariance model fits the data, then the researcher can continue to evaluate other models in the hierarchy. If a model with a higher level of invariance constraints does not fit worse than a model with a lower level of invariance constraints, then the researcher can conclude that measurement invariance is established at this higher level. The tests continue until the highest level of measurement invariance achieved is established for the measurement instrument under investigation in the data set at hand. However, if one of the models is rejected by the statistical tests, the tests do not provide information on the practical importance of the measurement non-invariance that is detected. Since dropping indicators that do not demonstrate measurement invariance will affect the content coverage of a measurement instrument, it is useful to have a method with which to gauge the practical significance of the differences in measurement model parameters over time (i.e. whether the differences have practical implications, Kirk, 1996). The primary concern of researchers conducting longitudinal studies is often whether the change in the observed responses to the indicators between measurement

occasions is due to true change in the mean/variance of the latent construct, or an artifact of the different values of the parameters in the measurement model across different measurement occasions. Drawing on studies of latent growth models and second-order latent growth models with continuous indicators (e.g.,Willson, 2014a; 2014b; Leite, 2007; Wirth, 2008), sensitivity analyses can be developed for ordered-categorical indicators that examine influences of longitudinal measurement non-invariance on estimates of growth parameters and their standard errors, assuming that the form of the latent growth model has been correctly specified.

The present study examined the suitability of using the changes in the second-order latent growth model parameters and in the corresponding standard errors as a potential sensitivity analysis to gauge the practical significance of longitudinal measurement non-invariance with ordered-categorical indicators. I first presented a brief review of longitudinal ordered-categorical CFA models, followed by a review of different levels of longitudinal measurement invariance and their interpretations. Then I discussed the importance of a sensitivity analysis to gauge the practical significance of longitudinal measurement non-invariance, and introduced the proposed sensitivity analysis based on the second-order latent growth model with ordered-categorical indicators. Following this, I described the design of this simulation study, the population model used to generate the data, and the evaluation criteria for results. Given the complex models in the simulation, I first examined the rates of model non-convergence and improper solution. Then I reported the results from the conditions with loading non-invariance, followed by the results from the conditions with threshold non-invariance, followed by the results from the conditions with unique factor non-invariance.

CHAPTER 2

LONGITUDINAL ORDERED-CATEGORICAL CFA MODEL

Let X_{ijt} represent the observed score from the i^{th} person on the j^{th} ordered-categorical indicator at measurement occasion t with score ranges $\{0, 1, ..., C\}$, where $c = 0, 1, \ldots, C$ are the response categories of the measured indicator. The CFA model for ordered-categorical indicators makes the assumption that there are continuous latent responses X^*_{ijt} (or underlying propensities, Bollen & Curran, 2006) that underlie each of the ordered-categorical observed responses X_{ijt}. The continuous latent responses are assumed to be multivariate normally distributed (Muthén, 1984), and they are sliced into the ordered-categorical observed responses by a set of threshold parameters ν for each indicator:

$$X_{ijt} = c, \text{ if } \nu_{jtc} \leq X^*_{ijt} < \nu_{jt(c+1)}, \tag{1}$$

where $c = 0, 1, \ldots, C$, the response categories of the ordered-categorical indicators, and $\{\nu_{jt0}, \nu_{jt1}, \ldots, \nu_{jt(C+1)}\}$ are the threshold parameters for indicator j at measurement occasion t ($\nu_{jt0} = -\infty$, and $\nu_{jt(C+1)} = \infty$). For any given latent response, the observed response is completely determined by the corresponding threshold parameters.

Assuming there is one latent common factor at each of the T measurement occasions[1], the longitudinal CFA model for the continuous latent responses is given by

$$X^*_{ijt} = \tau_{jt} + \lambda_{jt}\eta_{it} + u_{ijt}, \tag{2}$$

[1] Although the present work focuses on models with one latent common factor at each measurement occasion, it can be easily generalized to cases with more latent common factors per measurement occasion.

where τ_{jt} is the intercept, λ_{jt} is the factor loading of the continuous latent response *j* on the latent factor at measurement occasion *t*, η_{it} is the factor score for person *i* at measurement occasion *t*, and u_{ijt} is the unique factor score for person *i* on the j^{th} indicator at measurement occasion *t*. Typically, all intercepts τ_{jt} are constrained to zero to allow for the estimation of the latent threshold parameters.

To account for the longitudinal nature of the design, the common factors are allowed to freely correlate across time, with

$$\begin{aligned} \eta_{it} &\sim N(\boldsymbol{\kappa}, \boldsymbol{\Phi}), \\ \boldsymbol{\kappa} &= [\kappa_1, \kappa_2, \ldots, \kappa_T\ ]', \\ \boldsymbol{\Phi} &= \begin{bmatrix} \varphi_1 & \varphi_{12} & \cdots & \varphi_{1T} \\ \varphi_{21} & \varphi_2 & \cdots & \varphi_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_{T1} & \varphi_{T2} & \cdots & \varphi_T \end{bmatrix}. \end{aligned} \tag{3}$$

The diagonal elements of the common factor variance-covariance matrix $\boldsymbol{\Phi}$ represent the common factor variances at each occasion, and the off-diagonal elements of $\boldsymbol{\Phi}$ represent lagged common factor covariances across measurement occasions. With more measurement occasions and relatively few indicators per occasion, it may not be possible to freely estimate all lagged common factor covariances, otherwise the model may be underidentified. When this is the case, it may be reasonable to consider placing restrictions on the lagged common factor covariances, such as constraining covariances of the same lag to be equal (i.e. a Toeplitz structure, see Weiss, 2005), or constraining covariances of lag 2 and greater to zero. However, misspecifying the common factor variance-covariance matrix may lead to biased estimates of other model parameters, influencing the accuracy of tests of longitudinal measurement invariance. Thus, given

sufficient indicators per measurement occasion to ensure model identification, it is more appropriate to freely estimate all lagged common factor covariances.

In addition to allowing the common factors to freely correlate across measurement occasions, each unique factor is allowed to freely correlate with itself, but *not* with other unique factors, at other measurement occasions, with

$$u_{ijt} \sim N(0, \boldsymbol{\Theta}), \tag{4}$$

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\Theta}_{11} & \boldsymbol{\Theta}_{12} & \cdots & \boldsymbol{\Theta}_{1T} \\ \boldsymbol{\Theta}_{21} & \boldsymbol{\Theta}_{22} & \cdots & \boldsymbol{\Theta}_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\Theta}_{T1} & \boldsymbol{\Theta}_{T2} & \cdots & \boldsymbol{\Theta}_{TT} \end{bmatrix}.$$

$\boldsymbol{\Theta}$ is a super matrix, with each diagonal element $\boldsymbol{\Theta}_{tt}$ being a submatrix representing the unique factor variance-covariance matrix at measurement occasion *t*, and each off-diagonal element $\boldsymbol{\Theta}_{t,t+k}$ being a diagonal submatrix containing the lagged covariances of each unique factor with itself over time. Again, with more measurement occasions and relatively few indicators per occasion, it may not be possible to freely estimate all lagged unique factor covariances. In such cases, it may be reasonable to consider placing restrictions on $\boldsymbol{\Theta}_{t,t+k}$, such as constraining unique factor covariances of the same lag to be equal for each indicator separately (i.e. a Toeplitz structure), or constraining unique factor covariances of lag 2 or more to zero. However, given sufficient indicators per measurement occasion for model identification, it is more appropriate to freely estimate all lagged unique factor covariances. The basic model used to test longitudinal measurement invariance for ordered-categorical indicators is depicted in Figure 1.

CHAPTER 3

MODEL IDENTIFICATION FOR THE LONGITUDINAL ORDERED-CATEGORICAL CFA MODELS

In the CFA model for ordered-categorical indicators, neither the latent common factors nor the continuous latent responses have inherent scales (i.e., unit of measurement determining the mean and variance structures). Therefore, to identify the scales of the latent common factors and the continuous latent responses in the CFA model for ordered-categorical indicators, constraints must be imposed on the model parameters.

To identify the variance structure of a latent common factor, one of two strategies have commonly been employed with continuous measured variables: (1) The marker variable approach constrains the factor loading of one of the indicators (the marker variable) to 1.0, thereby giving the latent common factor a scale that is in the same unit as the marker variable; (2) a factor variance approach constrains the common factor variance to a fixed value, typically 1.0 (Bollen, 1989, p. 239). In the setting of longitudinal studies, it is common practice to use the marker variable approach to identify the common factor variance structure at all measurement occasions.

To identify the mean structure of a latent common factor with continuous measured variables, one of two strategies have commonly been employed with continuous measured variables: (1) Constrain the intercept of the marker variable to 0, or (2) constrain the common factor mean to 0. In longitudinal studies with latent common factors at multiple measurement occasions, various combinations of these two strategies may be used to identify the common factor mean structure. One common approach is to constrain the common factor mean to 0 at one measurement occasion (typically the first),

and to constrain the intercepts of the marker variables to be equal across time (e.g., Widaman et al., 2010).

Extending this work to ordered-categorical indicators with *three or more* response categories, there are several ways to identify the ordered-categorical CFA models involving multiple measurement occasions or multiple groups (e.g., Grimm, Ram, & Estabrook, in press; Millsap, 2011; Millsap & Tein, 2004). This study uses the following constraints (adapted from Millsap & Tein, 2004) to identify the mean and variance structures of both the latent common factors and the continuous latent responses in a longitudinal ordered-categorical CFA model:

1. The same observed indicator is chosen as the marker variable at all measurement occasions, with the factor loading constrained to 1.00.
2. The latent intercepts $\boldsymbol{\tau}_t$ are constrained to zero at all measurement occasions.
3. The common factor mean κ_t is constrained to zero at one measurement occasion (the reference measurement occasion, typically the first or last). At all other measurement occasions, the common factor mean is freely estimated.
4. The within-wave unique factor covariance matrix $\boldsymbol{\Theta}_{tt}$ is constrained to be an identity matrix ($\boldsymbol{\Theta}_{tt} = \boldsymbol{I}$) at the reference measurement occasion.[2] At all other measurement occasions, $\boldsymbol{\Theta}_{tt}$ is a diagonal matrix with the diagonal elements freely estimated.

[2] Alternatively, the total variances of all latent responses at the reference measurement occasion can be constrained to 1.0, instead of constraining the unique variances to 1.00 (adapted from Millsap & Tein, 2004).

5. One threshold for each indicator (and a *second* threshold for the marker variable) is constrained to be invariant across measurement occasions.

This model identification strategy makes it possible to freely estimate the unique factor variances at occasions other than the reference occasion while freely estimating the factor loadings and threshold parameters (other than the identification constraints), with indicators having *three or more* response categories. Thus, this identification strategy allows for the estimation of a configural invariance model (discussed below) that parallels the configural invariance model in the continuous case. With binary indicators, however, these constraints are not sufficient to identify the model. Since there is only one threshold per indicator, Constraint 5 above cannot be satisfied and additional constraints on other model parameters are needed to identify the means or variances of the continuous latent responses. For instance, additional constraints can be imposed on unique factor variances (or the total variances of continuous latent responses) at occasions other than the reference occasion (see also Koran & Hancock, 2010; Millsap & Tein, 2004). Alternatively, additional constraints could be imposed on factor loadings and the common factor means to identify the means of the continuous latent responses (see also Koran & Hancock, 2010; Grimm & Liu, in press). Thus with binary indicators, a configural invariance model that parallels the standard one for continuous indicators cannot be estimated. Moreover, since binary indicators have only one threshold per indicator, a test of threshold invariance cannot be achieved without other constraints to identify the means and variances of the latent responses (e.g., constraining all latent common factor means to zero and all unique factor variances to one at all occasions, Grimm et al, in press), which may be hard to meet in practice. Given these complications,

the current research will focus on ordered-categorical indicators with three or more response categories.

CHAPTER 4

TESTING LONGITUDINAL MEASUREMENT INVARIANCE WITH ORDERED-CATEGORICAL INDICATORS

Paralleling the set of hierarchical models compared in the test of measurement invariance with continuous indicators, to evaluate longitudinal measurement invariance with ordered-categorical indicators, a set of four hierarchical models can be sequentially tested and compared.

Model 1: The Configural Invariance Model

Paralleling the configural invariance model for continuous indicators, the configural invariance model for ordered-categorical indicators tests the hypothesis that the same general pattern of factor loadings holds across time. This model should provide a good fit to the data in order to continue evaluation of other models in the hierarchy.

Model 2: The Loading Invariance Model

Given a good fit of the configural invariance model to the data, the loading invariance model is fitted next, which adds the constraint that factor loadings are identical across measurement occasions: $\lambda_{11} = \lambda_{12} = \cdots = \lambda_{1T}$, $\lambda_{21} = \lambda_{22} = \cdots = \lambda_{2T}$, $\lambda_{31} = \lambda_{32} = \cdots = \lambda_{3T}$, …, with the first subscript *j* representing the indicator and the second subscript *t* representing time. The loading invariance model for ordered-categorical indicators parallels the weak invariance model for continuous indicators. Establishing longitudinal loading invariance for ordered-categorical indicators implies that changes over time in the expected *means* of the *continuous latent responses* can be fully explained by changes in the latent common factors over time (Appendix A). However, this condition is not sufficient to attribute changes over time in the expected *means* of the

observed responses solely to changes in the latent common factors: The continuous latent responses are inferred based on not only the observed responses but also distributional assumptions (multivariate normality) and threshold parameters.

Model 3: The Threshold Invariance Model

If the loading invariance model fits the data no worse than the configural invariance model, the threshold invariance model is then fitted, which adds the constraint that the threshold level of going from one response category to the next is equal across measurement occasions for each indicator: $\nu_{111} = \nu_{121} = \cdots = \nu_{1T1}$, $\nu_{112} = \nu_{122} = \cdots = \nu_{1T2}$, $\nu_{113} = \nu_{123} = \cdots = \nu_{1T3}$, …, with the first subscript *j* representing the indicator, the second subscript *t* representing time, and the third subscript *c* representing threshold. The threshold invariance model for ordered-categorical indicators parallels the strong invariance model for continuous indicators. However, unlike the case with continuous indicators, establishing loading and threshold invariance across measurement occasion does *not* imply that changes over time in the means of the *measured* ordered-categorical indicators can be entirely attributed to changes in the latent common factor. For that to be the case, the unique factor variances must also be invariant over time (see Appendix A for proof).

Model 4: The Unique Factor Invariance Model

If the threshold invariance model fits the data no worse than the loading invariance model, the unique factor invariance model is then fitted, which adds the constraint that the elements in $\boldsymbol{\Theta}_{tt}$ (all unique factor variances and any non-zero within-wave unique factor covariances) are equal across measurement occasions. On the other hand, the non-zero diagonal elements in $\boldsymbol{\Theta}_{t,t+k}$, the lagged unique factor covariances

across time, are freely estimated with no longitudinal equality constraints. Because unique variances at the reference occasion were fixed to 1.0 for identification purposes in earlier models, in the unique factor invariance model, all unique variances are fixed to 1.0.

The unique factor invariance model for ordered-categorical indicators parallels the strict invariance model for continuous indicators. Establishing longitudinal unique factor invariance implies that changes in the expected *means*, *variances*, and *within-wave covariances* of the *continuous latent responses* can be fully explained by changes in the latent common factors over time. More importantly, changes over time in the expected *means* and the *within-wave bivariate probabilities* of the ordered-categorical indicators can be fully explained by changes in the latent common factors (Appendix A).

CHAPTER 5

GAUGING THE PRACTICAL SIGNIFICANCE OF THE VIOLATIONS OF INVARIANCE

The need to achieve a more stringent level of measurement invariance (unique factor invariance) in order to compare the observed means of ordered-categorical indicators over time places a *stringent* requirement that will often not be met in practice. However, removing non-invariant indicators may impair the content validity of the measurement instrument. Thus, it can be helpful to conduct a sensitivity analysis that allows researchers to assess the practical significance of the failure to achieve a more advanced level in the hierarchy of levels of measurement invariance.

In longitudinal studies, the primary concern of researchers is often whether the change in the observed indicators between measurement occasions can be attributed to true change in the latent construct, or change in the psychometric properties of the measurement instrument. There have been some studies examining the influence of non-invariant continuous indicators on the parameter estimates (e.g., mean intercept and slope, intercept and slope variances and covariance) or the functional form of growth in a latent growth model or a second-order latent growth model. For instance, Leite (2007) compared parameter estimates from (a) a latent growth model based on item composites (means), (b) a latent growth model based on item composites (means) with fixed error variances estimated using the reliability of the composite, and (c) a second-order latent growth model with only identification constraints at the first-order level. Leite (2007) simulated a model with continuous indicators that achieved longitudinal configural invariance, weak invariance or strict invariance, and found that models (a) and (b)

produced biased parameter estimates when the indicators do not achieve strict invariance, whereas model (c) always produced adequate results. On the other hand, model (c) was found to be more likely to produce inadmissible solutions (i.e., Heywood cases), but this problem was alleviated by having more measurement occasions or larger sample sizes. Wirth (2008) compared the parameter estimates and the likelihood of accepting an alternative functional form in a latent growth model using composites of continuous indicators with those using factor scores saved from measurement models in which the factor loading of only one indicator or the factor loadings of all indicators were constrained to be equal over time. Wirth (2008) found that latent growth models using composites of continuous indicators or factor scores saved from measurement models with inappropriate invariance constraints tended to produce biased parameter estimates when the indicators were non-invariant. Model fit statistics from latent growth models using composites of continuous indicators had acceptable Type I error rates, and had increased likelihood of accepting an alternative form of growth as the level of loading non-invariance increased. On the other hand, model fit statistics from latent growth models using saved factor scores always had high Type I error rates and were biased towards accepting an alternative form of growth. In addition, also using continuous indicators, Kim and Willson (2014a, b) examined the influence of measurement *non-invariance across groups* on inferences of group differences in the mean intercept and mean linear slope parameters in latent growth modeling based on item composites or second-order latent growth models. These ideas can be generalized to ordered-categorical indicators: The biases of the growth parameter estimates in latent growth models due to longitudinal measurement non-invariance may be investigated as a potential sensitivity

analysis to gauge the practical significance of violations of longitudinal measurement invariance for ordered-categorical indicators.

Given that achieving unique factor invariance is theoretically required in order to compare the observed means of ordered-categorical indicators over time, latent growth models of the composite scores of the observed responses to the ordered-categorical indicators are *not* appropriate for the sensitivity analysis. Latent growth models of the saved factor scores are not appropriate for the sensitivity analysis either, provided the problems with the model fit statistics from such models with continuous indicators (Wirth, 2008), and the fact that regression coefficients are generally biased for both continuous and discrete indicators when treating the estimated latent common factor scores as observed (Hoshino & Bentler, 2011[3]). Instead, a second-order latent growth model can be fitted, with the first-order model being the measurement model for the ordered-categorical indicators, and the second-order model being a growth model. For instance, when the ordered-categorical indicators achieve longitudinal loading invariance but *not* threshold invariance, two second-order latent growth models can be fitted. The first model assumes (correctly) loading invariance at the first-order level; the second model assumes (incorrectly) threshold invariance at the first-order level. When the correct form of the latent growth model is specified at the second-order level, the discrepancies in the estimated growth parameters between these two models can be

[3] Hoshino and Bentler (2011) proposed an approach to reduce the bias resulting from using saved factor scores. However, this method sometimes produced biased estimates as compared to the generalized least squares methods using estimated polychoric or polyserial correlations (e.g., the weighted least squares methods implemented in M*plus*).

viewed as effect size estimates of the practical significance of the violations of longitudinal measurement invariance at the threshold invariance level.

CHAPTER 6

SPECIFICATION AND IDENTIFICATION OF THE SECOND-ORDER LATENT GROWTH MODEL

With the longitudinal CFA model for ordered-categorical indicators at the first-order level, a latent growth model can be fitted at the second-order level treating the latent common factors at each measurement occasion as outcomes of interest. The second-order latent growth model can be written as

$$\eta_i = \Gamma\xi_i + \zeta_i, \tag{5}$$

where η_i is a $T \times 1$ vector containing the first-order latent common factor scores for person *i* at all *T* measurement occasions, Γ is a $T \times R$ design matrix containing the factor loadings of the first-order latent common factor scores on the second-order latent growth factors ($R = 2$ for a linear growth model with a latent intercept factor and a latent linear slope factor), ξ_i is an $R \times 1$ vector containing the second-order latent growth factors, and ζ_i is a $T \times 1$ vector containing the disturbance scores (residuals) of first-order latent common factors for person *i* at all *T* measurement occasions (Figure 2). The second-order latent growth factors are typically assumed to be multivariate-normally distributed (Grimm et al., in press). For a linear latent growth model at the second-order level, when the initial measurement occasion is chosen as the reference occasion (*t* = 0), the $T \times R$ design matrix Γ is given by

$$\Gamma = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \dots & \dots \\ 1 & T-1 \end{bmatrix}. \tag{6}$$

The mean of the second-order latent intercept factor is constrained to zero to identify the model, whereas the mean of the second-order latent linear slope, the variances of the second-order growth factors and their covariance, are all freely estimated (Grimm et al., in press). The disturbance scores on each of the first-order latent common factors are assumed to follow a normal distribution with a mean of zero (Grimm et al., in press). Assuming that the correct form of the latent growth model is specified, covariances among the first-order latent common factors are typically assumed to be completely explained by the second-order growth factors, and the lagged disturbance covariances at the first-order level are typically constrained to zero (e.g., Grimm et al., in press). Note that this is different from the common practice in using CFA models to test longitudinal measurement invariance, where the latent common factors are allowed to freely correlate across measurement occasions. The specification of the first-order measurement model is the same as that of a longitudinal CFA model used to test longitudinal measurement invariance. In particular, given sufficient indicators per measurement occasion for model identification, all lagged unique factor covariances are freely estimated (e.g., Widaman et al., 2010).

The current study investigated the suitability of using the changes in the second-order latent growth model parameters and in the corresponding standard errors as a potential sensitivity analysis to gauge the practical significance of longitudinal measurement non-invariance with ordered-categorical indicators. Of central interest was how sensitive this sensitivity analysis was to different locations of non-invariance (i.e., non-invariance in the factor loadings, the threshold parameters, and the unique factor

variances). The current study also examined the influence of a number of other factors including the magnitude of non-invariance, the number of non-invariant indicators, the number of non-invariant occasions, and the number of response categories in the indicators.

CHAPTER 7

METHOD

A second-order latent growth model was used to generate the data, with a longitudinal ordered-categorical CFA model at the first-order level, and a latent linear growth model at the second-order level. The generated data consisted of four measurement occasions. For simplicity, the same set of five indicators were used to measure the same first-order common factor at each occasion with no missing data. This number of indicators per factor was in line with previous simulation studies and has been shown to produce accurate parameter estimates, particularly with a sufficient sample size (DiStefano & Morgan, 2014; Flora & Curran, 2004; Rhemtulla et al., 2012). A sample size of $N = 2000$ was used in the simulation study. Previous simulation studies suggest that various estimation methods for ordered-categorical CFA models such as Marginal Maximum Likelihood (MML), Diagonally Weighted Least Squares (DWLS), and Unweighted Least Squares (ULS) should provide accurate and similar results at such a sample size, especially with indicators that are *not* highly skewed (DiStefano & Morgan, 2014; Flora & Curran, 2004; Forero & Maydeu-Olivares, 2009; Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009; Yang-Wallentin, Jöreskog, & Luo, 2010). The current simulation study analyzed the generated data sets using the robust DWLS estimator in M*plus* (ESTIMATOR = WLSMV). This estimator provides more flexible scaling choices than MML. The Theta parameterization was used given the interest to evaluate the longitudinal invariance of unique factor variances[4].

[4] The Delta parameterization tends to generate more stable parameter estimates than the Theta parameterization in difficult conditions with small sample sizes, few (e.g., 3) indicators per latent common

In models with ordered-categorical indicators, when there is change over time in the mean level of the first-order common factor, it can happen that the response categories are well-populated at the first measurement wave, but show rather non-normal observed distributions and exhibit some categories with sparse data by the end of the longitudinal study, or vice versa. As a result, the bivariate or multivariate frequency table may have sparse or even empty cells, which can create problems for the estimation of polychoric correlations[5] (Brown & Bendetti, 1977; Flora & Curran, 2004; Bollen & Curran, 2006). These problems can potentially influence the parameter estimates from the second-order latent growth model of ordered-categorical indicators. As is explained in the later section "Population Model for Data Generation", population parameter values in this simulation study were chosen such that the lowest cell count in the *bivariate* frequency table at $N = 2000$ was expected to be around 5, to minimize the influence of sparse data while representing realistic research scenarios.

This simulation study used a 3 (Location of Measurement Non-Invariance: loading non-invariance only, threshold non-invariance only, or unique factor non-invariance only) × 2 (Magnitude of Non-Invariance for each non-invariant indicator: small versus large) × 2 (Number of Non-Invariant Indicators: one [X_{3t}] versus three [X_{3t}, X_{4t}, and X_{5t}]) × 2 (Number of Non-Invariant Occasions: one versus two) × 2 (Number of

factor, and highly skewed observed distributions of the indicators, particularly for binary indicators (Forero & Maydeu-Olivares, 2009; Muthén & Asparouhov, 2002). However, the two parameterizations have similar performance in less difficult conditions with larger sample sizes, more indicators per factor, and less skewed observed distributions of the indicators (Forero & Maydeu-Olivares, 2009). Moreover, the Delta parameterization does not permit direct specification of the unique factor invariance model.

[5] A polychoric correlation is the estimated theoretical correlation between two bivariate normal, continuous latent responses X^*_{ijt} based on the corresponding observed ordered-categorical responses.

Response Categories per indicator: three versus five) + 2 (baseline conditions with full measurement invariance: three versus five response categories) design to generate the data.

A total of 1000 replications were generated for each condition with non-invariant indicator(s) using M*plus* 7.11. Two different second-order latent linear growth models were then fitted to each generated data set in M*plus* 7.11. One model assumed the correct level of longitudinal measurement invariance at the first-order level, and the other model assumed an incorrect level of longitudinal measurement invariance one level higher in the hierarchy. For instance, when the data were simulated to have threshold non-invariance, I imposed constraints in line with longitudinal loading invariance (which was correct) in Analysis Model 1, and imposed constraints in line with longitudinal threshold invariance (which was *incorrect*) in Analysis Model 2. Comparing results from these two analysis models provided an estimate of the influence of assuming an incorrect level of longitudinal measurement invariance on inferences from second-order latent growth models.

To provide some idea about how much of the changes in parameter estimates between the two models with different invariance constraints was due to sampling variability, two baseline conditions with fully invariant indicators were also be simulated, one with three response categories per indicator, the other with five response categories per indicator. These baseline conditions had the same data generation models as those of the corresponding conditions with non-invariant indicators, except that the indicators were fully invariant over time in the baseline conditions. A total of 1000 replications were generated for each of the baseline conditions using M*plus* 7.11. Second-order latent

growth models were fitted to each generated data set of the baseline conditions in M*plus* 7.11. Four different levels of longitudinal measurement invariance constraints at the first-order level (configural invariance, loading invariance, threshold invariance, and unique factor invariance) were imposed. Any change in the second-order growth parameters, between models assuming different levels of measurement invariance fitted to the same simulated fully invariant data set using the same marker variable, should reflect sampling variability.

In all the analyses models, to achieve model identification, a marker variable strategy was used, constraining the factor loading of X_{1t}, an indicator that was always simulated to have full invariance, to 1.0 at all measurement occasions. The unique factor variances at the first measurement occasion was constrained to 1.0. The first threshold of all indicators and the second threshold of the marker variable X_{1t} were constrained to be equal across measurement occasions. The intercepts of all continuous latent responses and the intercepts of all first-order latent common factors were constrained to 0. The design matrix containing the factor loadings of the first-order latent common factor scores on the second-order latent growth factors were set to be the same as the one in the data generation model, and the mean of the second-order intercept factor was constrained to 0. The disturbance covariances of the first-order latent common factors were all constrained to 0, as in the data generation model. The details of each design factor in the current study are described below with a justification of the values selected for the study, followed by the details of the data generation model.

Details of Design Factors

Location of measurement non-invariance. Measurement non-invariance was simulated to occur in three different locations: factor loadings only, threshold parameters only, or unique factor variances only. Specifically, measurement non-invariance occurred at the last one or two of the four measurement occasions. For the factor loading non-invariance conditions, the factor loading(s) of the non-invariant indicator(s) at the non-invariant measurement occasion(s) were obtained by *subtracting* a constant from the corresponding factor loadings at earlier measurement occasions, as in previous simulations (e.g., Gonzalez-Roma, Hernandez, & Gomez-Benito, 2006; Kim & Yoon, 2011). For the threshold non-invariance conditions, the first two thresholds of the non-invariant indicator(s) were generated to be invariant over time, with the last threshold at the non-invariant measurement occasion(s) obtained by *adding* (e.g., Kim & Yoon, 2011; Stark, Chernyshenko, & Drasgow, 2006) a constant from the corresponding last thresholds at earlier measurement occasions[6]. For the unique factor non-invariance conditions, the unique factor variance(s) of the non-invariant indicator(s) at the non-invariant measurement occasion(s) were obtained by multiplying the corresponding unique factor variances at the earlier measurement occasions by a constant greater than 1.

Magnitude of non-invariance for each non-invariant indicator. The magnitude of non-invariance was set to be the same for each non-invariant indicator. For the factor loading non-invariance conditions, the small and large decreases in factor loadings corresponded to decreases on the metric of completely standardized factor loadings of .25

[6] Given the negatively skewed distributions of the ordered-categorical indictors at the last measurement occasion (see the later section "Population Model for Data Generation" for details), subtracting a constant from the last threshold may result in sparse or even empty cells.

and .50, respectively, relative to the first measurement occasion. These were in the range of the standardized values used in previous simulation studies for small/ large factor loading non-invariance (e.g., Gonzalez-Roma et al., 2006; Kim & Yoon, 2011; Meade & Lautenschlager, 2004). For the threshold non-invariance conditions, the small and large changes in thresholds corresponded to changes of .25 standard deviation and .50 standard deviation, respectively, of the continuous latent responses at the first measurement occasion. These values were in the range of the standardized values used in previous simulation studies for small/ large threshold non-invariance (e.g., Kim & Yoon, 2011). For the unique factor non-invariance conditions, the unique factor variances at earlier measurement occasions were multiplied by 1.5625 (= 1.25^2) and 2.25 (=1.5^2) to obtained the unique factor variance at the non-invariant measurement occasion(s), for small and large non-invariance conditions, respectively.

Number of non-invariant indicators. The number of non-invariant indicators, the number of occasions with measurement non-invariance, and the magnitude of non-invariance for each non-invariant indicator should all contribute to the total degree of measurement non-invariance in the model. Since the magnitude of non-invariance was set to be equal for each non-invariant indicator, with a certain magnitude of non-invariance per non-invariant indicator and a certain number of occasions with measurement non-invariance, the total degree of measurement non-invariance in the model should increase with the number of non-invariant indicators. The current simulation examined two levels of the number of non-invariant indicators: one versus three. When there was one non-invariant indicator, X_{3t} was simulated to be non-invariant. When there were three non-invariant indicators, X_{3t}, X_{4t}, and X_{5t} were simulated to be non-invariant.

Number of occasions with measurement non-invariance. Two conditions were simulated: one in which measurement non-invariance occurred at the last one of the four measurement occasions, and one in which measurement non-invariance occurred at the last two of the four measurement occasions.

Number of response categories per indicator. Two conditions were simulated: one in which all indicators had three response categories, and one in which all indicators had five response categories. At least three response categories would be needed to construct a configural invariance model paralleling the standard one for continuous indicators (as discussed in the previous section Model Identification for the Longitudinal Ordered-Categorical CFA Models). With more than five response categories, researcher may be more inclined to treat the indicators as continuous, especially when the observed distributions are non-normal leading to low cell counts in the bivariate frequency table, which may create estimation problems for ordered-categorical indicators. For instance, DiStefano and Morgan (2014) found that when the observed distribution is non-normal (*skewness* = 3, *kurtosis* = 7), the WLSMV estimator in M*plus* produced positively biased factor correlations, negatively biased standard errors of factor loadings, and negatively biased factor correlations for indicators with seven response categories even at $N = 800$, but accurate estimates for indicators with five response categories.

Population Model for Data Generation

For the first-order level measurement model, different levels of longitudinal measurement invariance were established over four measurement occasions. Non-invariance was generated at the last one or two measurement occasions. The location and magnitude of non-invariance were described above. The lagged unique factor correlations

followed a lag-1 autoregressive [AR(1)] structure, such that all unique factor correlations of lag-1 were set to $\rho_{jj(t,t+1)} = \rho$, all unique factor correlations of lag-2 were set to $\rho_{jj(t,t+2)} = \rho^2$, and all unique factor correlations of lag-3 were set to $\rho_{jj(t,t+3)} = \rho^3$. For the second-order latent growth model component, linear growth across equally spaced measurement occasions was simulated. The disturbances (residuals) of the first-order latent common factors were simulated to have zero correlation across time. The population parameter values were generated in two steps. In Step 1, parameter values were chosen to be in line with those used in previous simulation studies and were considered reasonable in real research. In Step 2, parameter values chosen in Step 1 were transformed to match the identification constraints in the analysis model (e.g., unique factor variances at Time 1 are equal to 1.0), such that the estimated parameter values could be compared to the population values directly.

Table 1 presents the population parameters selected in Step 1 for the baseline conditions with fully invariant indicators, before transformation. The factor loadings were chosen among values used in previous simulation studies of CFA or IRT models with ordered-categorical indicators (e.g., Kim & Yoon, 2011; Stark et al., 2006) and were expected to occur in real research settings (see DiStefano & Hess, 2005). The design matrix at the second-order level was chosen as follows to reflect a latent linear growth model with the first measurement occasion as the reference occasion:

$$\Gamma = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}. \quad (7)$$

The second-order intercept variance was chosen to be 0.5, and the second-order slope variance was chosen to be 0.1, one fifth of the intercept variance, following the suggestion by Muthén and Muthén (2002), as in previous simulation studies of latent growth models/ second-order latent growth models (Kim & Willson, 2014a; 2014b; Leite, 2007; Wirth, 2008). The intercept-slope covariance was chosen such that the intercept-slope correlation was .40 (Kim & Willson, 2014a; 2014b; Leite, 2007; Wirth, 2008). Values on the disturbance variances were chosen such that the first-order latent common factors had R^2 values of 0.70 (Wirth, 2008). Values on the unique factor variances were chosen to be 0.30, such that the continuous latent responses underlying the ordered-categorical indicators had R^2 values that range between 0.54 and 0.90, which were among the R^2 values used in previous simulation studies (e.g., Kim & Willson, 2014a; 2014b; Kim & Yoon, 2011; Wirth, 2008). Values on the thresholds for the five-category conditions were taken from a previous simulation study on multiple-group measurement non-invariance of ordered-categorical indicators (Kim & Yoon, 2011). The mean intercept was set to 0 (Kim & Willson, 2014a; 2014b; Leite, 2007). The mean slope was set to 0.6, such that the two lowest cell frequencies in the univariate frequency table with five response categories per indicator were 4.4% and 5.4% in a very large sample (N = 1,000,000). Thus at a sample size of N = 2000, the lowest cell count in the *univariate* frequency table with five response categories per indicator was expected to be around 80 (4.4% × 2000 = 88), and the lowest cell count in the *bivariate* frequency table with five response categories per indicator was expected to be around 5. Data generated by such population values made sure that most data fell in the lower response categories at the first measurement occasion, but fell in the higher response categories at the last

measurement occasion. Put differently, at the first measurement occasion the higher response categories were relatively sparse, but at the last measurement occasion the lower response categories were relatively sparse. Figures 3-7 present the distributions of the observed response categories at the first and the last measurement occasions for each indicator in the baseline condition with five response categories. The middle three response categories in the five-category conditions were collapsed to create data for the corresponding three-category conditions, such that the relatively sparse cells were maintained (different random seeds were used to generate the data, but the thresholds in the three-category conditions were the first and last thresholds in the corresponding five-category conditions).

Table 2 contains the population parameter values used for data generation in the baseline conditions with fully invariant indicators, after transformation in Step 2 to match identification constraints in the analysis models as described in the earlier section Model Identification for the Longitudinal Ordered-Categorical CFA Models.

Evaluation Criteria for Results

Given the complex models in the present study, the rates of model non-convergence and improper solution were examined to evaluate potential problems of estimation. Only converged solutions were considered for further analyses. Results from conditions with loading non-invariance are reported first, followed by results from conditions with threshold non-invariance, followed by results from conditions with unique factor non-invariance.

For conditions with measurement non-invariance, three evaluation criteria were considered: 1) the relative change[7] in the second-order latent growth parameters of interest and in the corresponding standard errors, between a correctly specified baseline model (e.g., assuming configural invariance) and an *incorrectly* specified more constrained model (e.g., assuming loading invariance); 2) the standardized change in the second-order mean linear slope, calculated as the ratio of the change in the estimated mean linear slope parameter over the square root of the estimated intercept variance from the correctly specified less constrained model[8]; and 3) the statistical power[9] of the nested model test to detect the incorrect measurement invariance constraints in the more constrained model assuming an incorrect level of measurement invariance.

In the baseline conditions with fully invariant indicators, I considered three corresponding evaluation criteria to provide benchmarks for comparison: 1) the relative change in each growth parameter of interest between a correctly specified baseline model assuming one level of measurement invariance (e.g., configural invariance) and a *correctly* specified more constrained model assuming a higher level of measurement invariance (e.g., loading invariance); 2) the standardized change in the second-order

[7] The relative change in a growth parameter was defined as the growth parameter estimate from a second-order latent growth model assuming the incorrect level of longitudinal measurement invariance at the lower level, minus the growth parameter from a second-order latent growth model assuming a less stringent, correct level of longitudinal measurement invariance, divided by the latter.

[8] The standardized change in the mean linear slope may be more informative than the relative change in the mean linear slope when the mean linear slope is close to zero.

[9] The statistical power is represented by the proportion of the 1000 replications for which a significant nested model test statistic is found in a condition with manipulated measurement non-invariance. For each location of measurement non-invariance, the design includes 2 (Magnitude of Non-Invariance) × 2 (Number of Non-Invariant Indicators) × 2 (Number of Non-Invariant Occasions) × 2 (Number of Response Categories) = 16 different conditions under which the nested model test can be performed.

mean linear slope between a correctly specified baseline model assuming one level of measurement invariance and a *correctly* specified more constrained model assuming a higher level of measurement invariance; and 3) the empirical Type 1 error rate[10] of the nested model test comparing the baseline model and a *correctly* specified more constrained model assuming a higher level of measurement invariance.

Three procedures were used to identify meaningful differences in the evaluation criteria as a function of the design factors in the conditions with measurement non-invariance. First, to visually portray the magnitude of the differences between the conditions, trellis plots were created across study conditions for each of the evaluation criteria. For evaluation criteria 1) and 2), the mean level of each evaluation criterion and the corresponding 95% normal-theory confidence interval[11] were graphed for each condition with manipulated measurement non-invariance. To provide a benchmark, the mean levels of these evaluation criteria and the 95% normal-theory confidence limits from the corresponding models in the corresponding baseline condition were also included in the trellis plots. According to Cumming and Finch (2005), when the 95% normal-theory confidence intervals do not overlap, the means of the two groups differ at

[10] The empirical Type 1 error rate is represented by the proportion of the 1000 replications for which a significant nested model test statistic is found in a baseline condition with fully invariant indicators. Since the nominated Type 1 error rate is .05, the standard error of the binomial distribution is $\sqrt{p(1-p)/n} = \sqrt{(.05)(.95)/1000} = .00689$. Thus, empirical Type 1 error rates that fell out of the range of the 95% confidence interval [.0365, .0635] were considered problematic.

[11] Most of the distributions of the evaluation criteria were closely approximated by the normal distribution. For those evaluation criteria with a rather non-normal (usually skewed) distribution, the use of the 95% normal-theory confidence limits, as compared to Tukey's box plot which provides a distribution-free representation, provided similar conclusions regarding whether the distribution of the evaluation criterion in a condition with measurement non-invariance was separated enough from the corresponding distribution in the baseline condition.

a level of significance that is at least $\alpha = .01$. Second, as a screening device, I identified those evaluation criteria for which the mean difference exceeded .10 between the baseline condition and at least one of the conditions with manipulated measurement non-invariance. These differences in the evaluation criteria were deemed to be of practical importance and worthy of further study. Finally, only for evaluation criteria that met the .10 difference standard, I conducted a between-subjects Analysis of Variance (ANOVA) for each location of non-invariance separately, to provide information about the importance of each of the factors in the design. For the relative changes in the growth parameters and the standardized change in the second-order mean linear slope, I conducted a 2 (Magnitude of Non-Invariance) × 2 (Number of Non-Invariant Indicators) × 2 (Number of Non-Invariant Occasions) × 2 (Number of Response Categories) between-subjects ANOVA. For the relative changes in the standard errors of the growth parameters, because the Number of Response Categories always had an impact in the baseline conditions with fully invariant indicators, I conducted a separate 2 (Magnitude of Non-Invariance) × 2 (Number of Non-Invariant Indicators) × 2 (Number of Non-Invariant Occasions) between-subjects ANOVA for conditions with three response categories and conditions with five response categories, respectively. With 1000 replications in each condition with manipulated measurement non-invariance, each location of non-invariance involved a large number of records to be submitted for analysis, resulting in very high power to detect very small effect sizes. Therefore η^2 was used as the effect size indicator, with $\eta^2 > .02$ used as the standard for effect sizes worthy of consideration. This magnitude is slightly above Cohen's (1988) value for a small effect size ($\eta^2 = .01$).

CHAPTER 8

RESULTS

Non-Convergence and Computational Problems

Overall, non-convergence only occurred for the second-order latent growth models assuming configural invariance. These models had the fewest constraints and the greatest number of parameters to estimate. Given the design of the study, the second-order latent growth models assuming configural invariance were estimated only in the baseline conditions with fully invariant indicators and the conditions with loading non-invariance.

Also in the baseline conditions with fully invariant indicators and the conditions with loading non-invariance, in a small proportion (< 1%) of the replications with no convergence problems, the DIFFTEST comparing the fit of the model assuming loading invariance versus the model assuming configural invariance could not be computed. When this happened, M*plus* generated the following error message: "THE CHI-SQUARE COMPUTATION COULD NOT BE COMPLETED BECAUSE OF A SINGULAR MATRIX."

Table 3 summarizes the rate of non-convergence and the rate of computational problems in the first 1000 replications for each condition in the baseline conditions with fully invariant indicators. Table 4 summarizes the rate of non-convergence and the rate of computational problems in the first 1000 replications for each condition in the conditions with loading non-invariance. Approximately 4 to 5% of the replications failed to produce a converged solution for the second-order latent growth model assuming configural invariance in the baseline conditions with fully invariant indicators. Approximately 4 to

8% of the replications failed to produce a converged solution for the second-order latent growth model assuming configural invariance in the conditions with loading non-invariance. I used 3000 iterations for each analysis model in the simulation, but providing start values based on population parameter values and increasing the number of iterations to 10,000 did not substantially reduce the model non-convergence rate in the replications that failed to produce a converged solution for the model assuming configural invariance. The rate of model non-convergence was slightly higher in the conditions with large loading non-invariance (lower part of Table 4) than in the conditions with small loading non-invariance (upper part of Table 4) or in the baseline conditions with fully invariant indicators (Table 3). This was probably due to the fact that indicators with non-invariant factor loading(s) were simulated to have lower factor loadings at the last one or two occasions, and indicators at the last measurement occasion had some relatively sparse response categories by design (see Figures 3-7). In the conditions with large loading non-invariance, the population values of the non-invariant factor loadings at the last one or two occasions (range from .22 to .35) were even lower than in the conditions with small loading non-invariance (range from .46 to .62). In the baseline conditions with fully invariant indicators and the conditions with loading non-invariance, the rate of computational problems was always less than 1%. The replications with either convergence problems or computational problems were replaced by additional replications that had converged solutions with no computational problem, so that each condition in this study had 1000 replications with converged solutions with no computational problem for further analyses.

Conditions with Loading Non-Invariance

Relative changes in the second-order growth parameters. Given that the mean intercept was constrained to 0 for model identification, I report the relative changes in four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance.

Mean linear slope. I calculated the relative change (*RC*) in the estimated second-order mean linear slope ($\overline{RC}_{\text{mean slope}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 8 shows the $\overline{RC}_{\text{mean slope}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{mean slope, three response categories}} = .000$; $\overline{RC}_{\text{mean slope, five response categories}} = .003$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{mean slope}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 8), suggesting that these $\overline{RC}_{\text{mean slope}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{mean slope}}$. The difference in $\overline{RC}_{\text{mean slope}}$ exceeded .10 between the baseline condition and all four conditions with large loading non-invariance for three indicators and two

conditions with small loading non-invariance for three indicators (lower panels of Figure 8). Thus, a between-subjects ANOVA was conducted on $RC_{\text{mean slope}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .460$): On average $\overline{RC}_{\text{mean slope}}$ was close to zero when there was one non-invariant indicator ($\overline{RC}_{\text{mean slope, one non-invariant indicator}} = -.005$), but became more negative when there were three non-invariant indicators ($\overline{RC}_{\text{mean slope, three non-invariant indicators}} = -.172$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .104$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero with small non-invariance ($\overline{RC}_{\text{mean slope, small non-invariance}} = -.049$) than with large non-invariance ($\overline{RC}_{\text{mean slope, large non-invariance}} = -.128$). The Number of Response Categories in the indicators showed a main effect ($\eta^2 = .075$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero when the indicators had three response categories ($\overline{RC}_{\text{mean slope, three response categories}} = -.055$) than when the indicators had five response categories ($\overline{RC}_{\text{mean slope, five response categories}} = -.122$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .040$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero with one non-invariant occasion ($\overline{RC}_{\text{mean slope, one non-invariant occasion}} = -.064$) than with two non-invariant occasions ($\overline{RC}_{\text{mean slope, two non-invariant occasions}} = -.113$).

These main effects were modified by three two-way interactions: A Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .092$), a

Number of Non-Invariant Indicators by Number of Response Categories interaction ($\eta^2 = .067$), and a Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .032$). As shown in Figure 8, when there was one non-invariant indicator, the influence of the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories was negligible. In contrast, when there were three non-invariant indicators, the influence of the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories became much larger. The $\overline{RC}_{\text{mean slope}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions, especially when the indicators had five instead of three response categories.

Intercept variance. I calculated the relative change (*RC*) in the estimated second-order intercept variance ($\overline{RC}_{\text{intercept variance}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 9 shows the $\overline{RC}_{\text{intercept variance}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{intercept variance, three response categories}} = .001$; $\overline{RC}_{\text{intercept variance, five response categories}} = .001$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{intercept variance}}$ in the baseline conditions. Several conditions with three non-invariant

indicators had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept variance}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 9), suggesting that these $\overline{RC}_{\text{intercept variance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was *not* met for $\overline{RC}_{\text{intercept variance}}$. The greatest difference in $\overline{RC}_{\text{intercept variance}}$ between the baseline condition and a condition with manipulated loading non-invariance was .088. Thus, a between-subjects ANOVA was *not* conducted on $RC_{\text{intercept variance}}$.

Linear slope variance. I calculated the relative change (*RC*) in the estimated second-order linear slope variance ($\overline{RC}_{\text{slope variance}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 10 shows the $\overline{RC}_{\text{slope variance}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{slope variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{slope variance, three response categories}} = .013$; $\overline{RC}_{\text{slope variance, five response categories}} = .032$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{slope variance}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{\text{slope variance}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 10), suggesting that these

$\overline{RC}_{\text{slope variance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{slope variance}}$. The difference in $\overline{RC}_{\text{slope variance}}$ exceeded .10 between the baseline condition and all conditions with three non-invariant indicators (lower panels of Figure 10). Thus, a between-subjects ANOVA was conducted on $RC_{\text{slope variance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .593$): On average $\overline{RC}_{\text{slope variance}}$ was close to zero when there was one non-invariant indicator ($\overline{RC}_{\text{slope variance, one non-invariant indicator}} = .010$), but became more negative when there were three non-invariant indicators ($\overline{RC}_{\text{slope variance, three non-invariant indicators}} = -.429$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .067$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero with small non-invariance ($\overline{RC}_{\text{slope variance, small non-invariance}} = -.136$) than with large non-invariance ($\overline{RC}_{\text{slope variance, large non-invariance}} = -.284$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .035$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero with one non-invariant occasion ($\overline{RC}_{\text{slope variance, one non-invariant occasion}} = -.157$) than with two non-invariant occasions ($\overline{RC}_{\text{slope variance, two non-invariant occasions}} = -.263$).

These main effects were modified by a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .064$) and a Number of Non-Invariant

Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .033$). As shown in Figure 10, when there was one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, when there were three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{\text{slope variance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions.

Intercept-slope covariance. I calculated the relative change (*RC*) in the estimated second-order intercept-slope variance ($\overline{RC}_{\text{intercept-slope covariance}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 11 shows the $\overline{RC}_{\text{intercept-slope covariance}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept-slope covariance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{slope variance, three response categories}} = .013$; $\overline{RC}_{\text{slope variance, five response categories}} = .032$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{intercept-slope covariance}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept-slope covariance}}$ that did not overlap with that in the corresponding baseline condition (lower panels of

Figure 11), suggesting that these $\overline{RC}_{\text{intercept-slope covariance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{intercept-slope covariance}}$. The difference in $\overline{RC}_{\text{intercept-slope covariance}}$ exceeded .10 between the baseline condition and all conditions with three non-invariant indicators (lower panels of Figure 11). Thus, a between-subjects ANOVA was conducted on $RC_{\text{intercept-slope covariance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .569$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was close to zero when there was one non-invariant indicator ($\overline{RC}_{\text{intercept-slope covariance, one non-invariant indicator}} = .019$), but became negative when there were three non-invariant indicators ($\overline{RC}_{\text{intercept-slope covariance, three non-invariant indicators}} = -.606$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .106$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero with small non-invariance ($\overline{RC}_{\text{intercept-slope covariance, small non-invariance}} = -.159$) than with large non-invariance ($\overline{RC}_{\text{intercept-slope covariance, large non-invariance}} = -.429$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .031$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero with one non-invariant occasion ($\overline{RC}_{\text{intercept-slope covariance, one non-invariant occasion}} = -.221$) than with two non-invariant occasions ($\overline{RC}_{\text{intercept-slope covariance, two non-invariant occasions}} = -.366$).

These main effects were modified by a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .105$) and a Number of Non-Invariant

Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .029$). As shown in Figure 11, when there was one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, when there were three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{\text{intercept-slope covariance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions.

Relative changes in the standard errors of the second-order growth parameters. Because the mean intercept was constrained to 0 for model identification and the corresponding standard error was 0, I report the relative changes in the standard errors of four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance. I report the relative changes in the standard errors of these growth parameters for completeness. For those second-order growth parameters for which the mean difference in the relative changes exceeded .10 (indicating material bias) between the baseline condition and at least one of the conditions with loading non-invariance, standard errors are clearly of only secondary interest, and thus the corresponding ANOVA results were not reported.

***Standard error of the mean linear slope*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order mean linear slope ($\overline{RC}_{SE_{\text{mean slope}}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 12

shows the $\overline{RC}_{SE_{\text{mean slope}}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{mean slope}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{mean slope}},\text{ three response categories}} = .046$; $\overline{RC}_{SE_{\text{mean slope}},\text{ five response categories}} = -.181$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{mean slope}}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{mean slope}}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 12), suggesting that these $\overline{RC}_{SE_{\text{mean slope}}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{mean slope}}}$. The difference in $\overline{RC}_{SE_{\text{mean slope}}}$ exceeded .10 between the baseline condition and several conditions with three non-invariant indicators (lower panels of Figure 12). When there were three response categories in the indicators as shown in the left panels of Figure 12, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{SE_{\text{mean slope}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. When there were five response categories in the indicators

as shown in the right panels of Figure 12, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. Again the $\overline{RC}_{SE_{\text{mean slope}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the mean linear slope.

Standard error of the intercept variance. I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept variance ($\overline{RC}_{SE_{\text{intercept variance}}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 13 shows the $\overline{RC}_{SE_{\text{intercept variance}}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept variance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three response categories}} = -.015$; $\overline{RC}_{SE_{\text{intercept variance}},\text{ five response categories}} = -.012$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept variance}}}$ in the baseline conditions. Two conditions with three non-invariant indicators and three response categories had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept variance}}}$ that did not overlap with that in the corresponding baseline condition (lower-left panel of Figure 13), suggesting that these

$\overline{RC}_{SE_{\text{intercept variance}}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{intercept variance}}}$. The difference in $\overline{RC}_{SE_{\text{intercept variance}}}$ exceeded .10 between the baseline condition and one condition with loading non-invariance (large non-invariance for three indicators at two occasions with three response categories; lower-left panel of Figure 13). Note that the .10 difference standard for a meaningful difference was *not* met for the relative changes in the corresponding growth parameter, the intercept variance, and a between-subjects ANOVA was *not* conducted on the relative changes in the intercept variance. Thus, the ANOVA results on the relative changes in the corresponding *standard error* ($RC_{SE_{\text{intercept variance}}}$) are reported here. Because the Number of Response Categories had an impact on $RC_{SE_{\text{intercept variance}}}$ in the baseline conditions with fully invariant indicators, a separate between-subjects ANOVA was conducted on $RC_{SE_{\text{intercept variance}}}$ for the loading non-invariance conditions with three response categories and those with five response categories, respectively.

For conditions with loading non-invariance and indicators with three response categories (left panels in Figure 13), the ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .458$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was negative and closer to the corresponding value in the baseline condition (-.015) when there was one non-invariant indicator ($\overline{RC}_{SE_{\text{intercept variance, one non-invariant indicator}}} = -.013$), but became positive and farther away from the

corresponding value in the baseline condition when there were three non-invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three non-invariant indicators}} = .046$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .163$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was negative and closer to the corresponding value in the baseline condition with small non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\text{ small non-invariance}} = -.001$), but became positive and farther away from the corresponding value in the baseline condition with large non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\text{ large non-invariance}} = .034$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .058$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was closer to the corresponding value in the baseline condition with one non-invariant occasion ($\overline{RC}_{SE_{\text{intercept variance}},\text{ one non-invariant occasion}} = .006$) than with two non-invariant occasions ($\overline{RC}_{SE_{\text{intercept variance}},\text{ two non-invariant occasions}} = .027$).

These main effects were modified by three two-way interactions and a three-way interaction: A Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .139$), a Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .044$), a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .032$), and a Number of Non-Invariant Indicators by Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .025$). As shown in the left panels of Figure 13, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, when there were three non-invariant indicators, the influence of the Number of Non-Invariant Occasions

was almost negligible with small non-invariance, but became much larger with large non-invariance. The $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the pattern of results for the relative changes in the intercept variance in the conditions with three response categories in the indicators.

For conditions with loading non-invariance and indicators with five response categories (right panels in Figure 13), the ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .391$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was less negative and closer to the corresponding value in the baseline condition (-.012) when there was one non-invariant indicator ($\overline{RC}_{SE_{\text{interccept variance}},\ \text{one non-invariant indicator}} = -.013$), but became more negative and farther away from the corresponding value in the baseline condition when there were three non-invariant indicators ($\overline{RC}_{SE_{\text{interccept variance}},\ \text{three non-invariant indicators}} = -.050$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .063$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was less negative and closer to the corresponding value in the baseline condition with small non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\ \text{small non-invariance}} = -.024$) than with large non-invariance ($\overline{RC}_{SE_{\text{interceptvariance}},\ \text{large non-invariance}} = -.039$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .059$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was less negative and closer to the corresponding value in the baseline condition with one non-invariant occasion ($\overline{RC}_{SE_{\text{intercept variance}},\ \text{one non-invariant occasion}} = -.024$) than with two non-invariant occasions

($\overline{RC}_{SE_{\text{intercept variance}},\text{ two non-invariant occasions}} = -.039$). These main effects were modified by a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .062$) and a Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .058$). As shown in the right panels of Figure 13, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became larger. The $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. Note however that even when the $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition, its difference from the baseline value was .072, which was less than .10, and its 95% normal-theory confidence interval overlapped with that in the corresponding baseline condition. This pattern of results paralleled the pattern of results for the relative changes in the intercept variance in the conditions with five response categories in the indicators.

Standard error of the linear slope variance. I calculated the relative change (*RC*) in the estimated standard error of the second-order linear slope variance ($\overline{RC}_{SE_{\text{slope variance}}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 14 shows the $\overline{RC}_{SE_{\text{slope variance}}}$ value with the 95% normal-theory confidence limits for each

condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{slope variance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{slope variance}},\text{ three response categories}} = -.044$; $\overline{RC}_{SE_{\text{slope variance}},\text{ five response categories}} = -.276$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{slope variance}}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{slope variance}}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 14), suggesting that these $\overline{RC}_{SE_{\text{slope variance}}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{slope variance}}}$. The difference in $\overline{RC}_{SE_{\text{slope variance}}}$ exceeded .10 between the baseline condition and several conditions with three non-invariant indicators (lower panels of Figure 14). When there were three response categories in the indicators as shown in the left panels of Figure 14, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{SE_{\text{slope variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. When there were five response categories in the indicators as shown in the right panels of Figure 14, with one non-invariant indicator, the influence

of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{SE_{\text{slope variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the linear slope variance.

***Standard error of the intercept-slope covariance*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept-slope covariance ($\overline{RC}_{SE_{\text{intercept-slope covariance}}}$) in the conditions with loading non-invariance, comparing the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 15 shows the $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value with the 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{intercept-slope covariance, three response categories}}} = -.033$; $\overline{RC}_{SE_{\text{intercept-slope covariance, five response categories}}} = -.208$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 15),

suggesting that these $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$. The difference in $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ exceeded .10 between the baseline condition and all four conditions with large non-invariance on three indicators and three conditions with small non-invariance on three indicators (lower panels of Figure 15). When there were three response categories in the indicators as shown in the left panels of Figure 15, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance was negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance became much larger. The $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. When there were five response categories as shown in the right panels of Figure 15, with one non-invariant indicator, the influence of the Magnitude of Non-Invariance was negligible. In contrast, with three non-invariant indicators, the influence of the Magnitude of Non-Invariance became larger. The $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the intercept-slope covariance.

Standardized change in the second-order mean linear slope. I calculated the standardized change (*STDC*) in the estimated second-order mean linear slope ($\overline{STDC}_{\text{mean slope}}$) in the conditions with loading non-invariance, comparing the model

correctly assuming configural invariance to the model incorrectly assuming loading invariance. Figure 16 shows the $\overline{STDC}_{\text{mean slope}}$ value with 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{STDC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{STDC}_{\text{mean slope, three response categories}} = .000$; $\overline{STDC}_{\text{mean slope, five response categories}} = .001$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{STDC}_{\text{mean slope}}$ in the baseline conditions. Several conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{STDC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 16), suggesting that these $\overline{STDC}_{\text{mean slope}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{STDC}_{\text{mean slope}}$. The difference in $\overline{STDC}_{\text{mean slope}}$ exceeded .10 between the baseline condition and three of the conditions with large loading non-invariance for three indicators, as well as one condition with small loading non-invariance for three indicators with five response categories (lower panels of Figure 16). Thus, a between-subjects ANOVA was conducted on $STDC_{\text{mean slope}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .450$): On average $\overline{STDC}_{\text{mean slope}}$ was close to zero when

there was one non-invariant indicator ($\overline{STDC}_{\text{mean slope, one non-invariant indicator}} = -.005$), but became more negative when there were three non-invariant indicators ($\overline{STDC}_{\text{mean slope, three non-invariant indicators}} = -.147$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .101$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero with small non-invariance ($\overline{STDC}_{\text{mean slope, small non-invariance}} = -.042$) than with large non-invariance ($\overline{STDC}_{\text{mean slope, large non-invariance}} = -.110$). The Number of Response Categories in the indicators showed a main effect ($\eta^2 = .077$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero when the indicators had three response categories ($\overline{STDC}_{\text{mean slope, three response categories}} = -.047$) than when the indicators had five response categories ($\overline{STDC}_{\text{mean slope, five response categories}} = -.105$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .039$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero with one non-invariant occasion ($\overline{STDC}_{\text{mean slope, one non-invariant occasion}} = -.055$) than with two non-invariant occasions ($\overline{STDC}_{\text{mean slope, two non-invariant occasions}} = -.097$).

These main effects were modified by three two-way interactions: A Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .091$), a Number of Non-Invariant Indicators by Number of Response Categories interaction ($\eta^2 = .065$), and a Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .031$). As shown in Figure 16, when there was one non-invariant indicator, the influence of the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories was negligible. In

contrast, when there were three non-invariant indicators, the influence of the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories became much larger. The $\overline{STDC}_{\text{mean slope}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions, especially when the indicators had five instead of three response categories. This pattern of results paralleled the results for the relative changes in the mean linear slope.

Statistical power of the nested model test to detect the incorrect loading invariance constraints. In the baseline conditions, I computed the empirical Type 1 error rates, i.e. the proportion of the 1000 replications for which a significant test statistic was found, for the nested model test[12] comparing the model fit of (a) the second-order latent growth model correctly assuming configural invariance, and (b) the second-order latent growth model *correctly* assuming loading invariance. The empirical Type 1 error rate was .055 when the indicators had three response categories, and .038 when the indicators had five response categories. Since these values were both within the acceptable range of [.0365, .0635], I concluded that this nested model test was not biased in terms of the Type 1 error rate. I then calculated the statistical power of this nested model test in the conditions with manipulated loading non-invariance, examining the difference in model fit between (a) the second-order latent growth model correctly assuming configural invariance, and (b) the second-order latent growth model *incorrectly* assuming loading invariance. Figure 17 shows the statistical power of the nested model test for each

[12] This nested model test was performed using the DIFFTEST command in M*plus*.

condition with loading non-invariance. As can be seen on the figure, when there was one indicator with non-invariant factor loadings, the nested model test had relatively low statistical power (< .15) to detect the incorrect loading invariance constraints. Low statistical power characterized each of the different combinations of the factors of Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories in the indicators. In contrast, when there were three indicators with non-invariant factor loadings, the nested model test had very high statistical power (between .978 and 1.00) to detect the incorrect loading invariance constraints across different combinations of the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories in the indicators.

Summary. In the conditions with loading non-invariance, the Number of Non-Invariant Indicators showed a substantial influence on all components (growth parameters and standard errors) of evaluation criteria 1) and 2) except for the relative change in the intercept variance, for which no meaningful difference was found between the baseline condition and the conditions with loading non-invariance. For all other components of the evaluation criteria 1) and 2), when there was one indicator with non-invariant factor loadings across time, the influence of the other design factors (the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories) was always negligible. In contrast, when there were three indicators with non-invariant factor loadings across time, the influence of the other design factors became much larger. With three non-invariant indicators, the greater the Magnitude of Non-Invariance, the farther away the average evaluation criterion value was from the value in the corresponding baseline condition. The greater the Number of Non-Invariant

Occasions, the farther away the average evaluation criterion value was from the value in the corresponding baseline condition.

It is also noteworthy that the Number of Response Categories in the indicators had an influence on evaluation criteria 1) and 2). The difference in the average evaluation criterion value from the corresponding baseline value tended to be greater when the indicators had five response categories rather than three response categories, especially for the relative changes in the growth parameters and the standardized change in the mean linear slope. The average evaluation criterion value was farthest away from the value in the corresponding baseline condition when there was large loading non-invariance for three indicators at two measurement occasions with five response categories.

In addition, when the indicators had five response categories, the distributions of the evaluation criteria were always much wider than when the indicators had three response categories, and this was true for both the conditions with non-invariant loadings and the baseline conditions with fully invariant indicators. This result may be related to the increased number of parameters that needed to be estimated in the second-order latent growth models assuming longitudinal configural and loading invariance when there were five rather than three response categories in the indicators.

Also, when the indicators had five response categories, the standard errors of the growth parameters always decreased substantially as a result of adding loading invariance constraints; when the added loading invariance constraints were incorrect (i.e., in the loading non-invariance conditions), the decrease was greater than when the loading invariance constraints were correct (i.e., in the baseline conditions with fully invariant

indicators). One implication is that the conclusion of statistical significance of a growth parameter of interest may change as more invariance constraints are added, whether or not the added invariance constraints are appropriate. In contrast, when the indicators had three response categories, the standard errors of the growth parameters did *not* change substantially as a result of adding *correct* loading invariance constraints in the baseline conditions. These standard errors sometimes changed substantially as a result of adding *incorrect* loading invariance constraints in the loading non-invariance conditions: The change tended to be positive for the standard error of the intercept variance, but negative for the standard errors of the other growth parameters.

The nested model likelihood ratio test comparing the fit of the second-order latent growth model assuming configural invariance with that of the model assuming loading invariance had acceptable Type 1 error rates. This nested model test had very high statistical power (above .95) to detect loading non-invariance for three indicators, but had very low statistical power (below .15) to detect loading non-invariance for one indicator.

Conditions with Threshold Non-Invariance

Relative changes in the second-order growth parameters. Again, since the mean intercept was constrained to 0 for model identification, I report the relative changes in four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance.

***Mean linear slope*.** I calculated the relative change (*RC*) in the estimated second-order mean linear slope ($\overline{RC}_{\text{mean slope}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 18 shows the $\overline{RC}_{\text{mean slope}}$ value with 95% normal-

theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{mean slope, three response categories}} = .000$; $\overline{RC}_{\text{mean slope, five response categories}} = .000$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{mean slope}}$ in the baseline conditions. Three conditions with three non-invariant indicators and one condition with one non-invariant indicator, all with three response categories in the indicators, had a 95% normal-theory confidence interval of $\overline{RC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline condition (left panels of Figure 18), suggesting that these $\overline{RC}_{\text{mean slope}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{mean slope}}$. The difference in $\overline{RC}_{\text{mean slope}}$ exceeded .10 between the baseline condition and one condition with large threshold non-invariance for three indicators with three response categories at two occasions (lower-left panel of Figure 18). Thus, a between-subjects ANOVA was conducted on $RC_{\text{mean slope}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Occasions had a main effect ($\eta^2 = .169$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero (which was also the value in the corresponding baseline condition) with one non-invariant occasion ($\overline{RC}_{\text{mean slope, one non-invariant occasion}} = -.027$) than with two non-invariant occasions

($\overline{RC}_{\text{mean slope, two non-invariant occasions}} = -.071$). The Number of Non-Invariant Indicators showed a main effect ($\eta^2 = .163$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero (which was also the value in the corresponding baseline condition) when there was one non-invariant indicator ($\overline{RC}_{\text{mean slope, one non-invariant indicator}} = -.028$) than when there were three non-invariant indicators ($\overline{RC}_{\text{mean slope, three non-invariant indicators}} = -.071$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .115$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero (which was also the value in the corresponding baseline condition) with small non-invariance ($\overline{RC}_{\text{mean slope, small non-invariance}} = -.031$) than with large non-invariance ($\overline{RC}_{\text{mean slope, large non-invariance}} = -.068$). The Number of Response Categories in the indicators also showed a main effect ($\eta^2 = .115$): On average $\overline{RC}_{\text{mean slope}}$ was closer to zero when the indicators had five response categories ($\overline{RC}_{\text{mean slope, five response categories}} = -.031$) than when the indicators had three response categories ($\overline{RC}_{\text{mean slope, three response categories}} = -.068$).

These main effects were modified by four two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .034$), a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .023$), a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .022$), and a Number of Non-Invariant Indicators by Number of Response Categories interaction ($\eta^2 = .021$). As shown in Figure 18, in general, there was a multiplicative effect of the design factors. When there was one non-invariant indicator, the differences in $\overline{RC}_{\text{mean slope}}$ from the corresponding baseline values were all less

than .10 (the standard for a meaningful difference), although a small multiplicative effect of the other factors could be observed. When there were three non-invariant indicators, the multiplicative effect of the other design factor became much larger: $\overline{RC}_{\text{mean slope}}$ increased when the Magnitude of Non-Invariance was large rather than small and there were two non-invariant occasions rather than one. In contrast to the earlier results in the loading non-invariance conditions, the use of five rather than three response categories decreased the amount of $\overline{RC}_{\text{mean slope}}$. Interestingly, in all cases in which there was either (a) one non-invariant indicator, (b) one non-invariant occasion, or (c) five response categories, the difference in $\overline{RC}_{\text{mean slope}}$ from the corresponding baseline condition did not exceed the .10 standard. The $\overline{RC}_{\text{mean slope}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions with three response categories.

Intercept variance. I calculated the relative change (*RC*) in the estimated second-order intercept variance ($\overline{RC}_{\text{intercept variance}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 19 shows the $\overline{RC}_{\text{intercept variance}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{intercept variance, three response categories}} = .000$; $\overline{RC}_{\text{intercept variance, five response categories}} = -.001$), and the

dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{intercept variance}}$ in the baseline conditions. Four conditions with three non-invariant indicators and one condition with one non-invariant indicator had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept variance}}$ that did not overlap with that in the corresponding baseline condition, suggesting that these $\overline{RC}_{\text{intercept variance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was *not* met for $\overline{RC}_{\text{mean slope}}$. Note that only one condition with non-invariant thresholds showed a difference in $\overline{RC}_{\text{intercept variance}}$ from the corresponding baseline condition that is approaching .10 ($\overline{RC}_{\text{intercept variance}} = .099$, large non-invariance on three non-invariant indicators at two occasions with five response categories). All other conditions had a difference in $\overline{RC}_{\text{intercept variance}}$ from the corresponding baseline condition that was less than .10. Thus, a between-subjects ANOVA was *not* conducted on $RC_{\text{intercept variance}}$.

Linear slope variance. I calculated the relative change (*RC*) in the estimated second-order linear slope variance ($\overline{RC}_{\text{slope variance}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 20 shows the $\overline{RC}_{\text{slope variance}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{slope variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{slope variance, three response categories}} = .000$; $\overline{RC}_{\text{slope variance, five response categories}} = .001$), and the dotted

lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{slope variance}}$ in the baseline conditions. One condition with threshold non-invariance had a 95% normal-theory confidence interval of $\overline{RC}_{\text{slope variance}}$ that did not overlap with that in the corresponding baseline condition (large non-invariance on three non-invariant indicators at two occasions with three response categories), suggesting that this $\overline{RC}_{\text{slope variance}}$ value differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{slope variance}}$. The difference in $\overline{RC}_{\text{slope variance}}$ exceeded .10 between the baseline condition and several conditions with two non-invariant occasions. Thus, a between-subjects ANOVA was conducted on $RC_{\text{slope variance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Occasions had a main effect ($\eta^2 = .259$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition with one non-invariant occasion ($\overline{RC}_{\text{slope variance, one non-invariant occasion}} = -.018$) than with two non-invariant occasions ($\overline{RC}_{\text{slope variance, two non-invariant occasions}} = -.121$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .078$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition with small non-invariance ($\overline{RC}_{\text{slope variance, small non-invariance}} = -.041$) than with large non-invariance ($\overline{RC}_{\text{slope variance, large non-invariance}} = -.097$). The Number of Non-Invariant Indicators showed a

main effect ($\eta^2 = .045$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{RC}_{\text{slope variance, one non-invariant indicator}} = -.048$) than when there were three non-invariant indicators ($\overline{RC}_{\text{slope variance, three non-invariant indicators}} = -.091$). The Number of Response Categories in the indicators also showed a main effect ($\eta^2 = .039$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition when the indicators had five response categories ($\overline{RC}_{\text{slope variance, five response categories}} = -.049$) than when the indicators had three response categories ($\overline{RC}_{\text{slope variance, three response categories}} = -.089$).

These main effects were modified by two two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .036$) and a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .035$). As shown in Figure 20, when there was one non-invariant occasion, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators were both negligible. In contrast, when there were two non-invariant occasions, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{\text{slope variance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two occasions with three response categories.

Intercept-slope covariance. I calculated the relative change (*RC*) in the estimated second-order intercept-slope variance ($\overline{RC}_{\text{intercept-slope covariance}}$) in the conditions with

threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 21 shows the $\overline{RC}_{\text{intercept-slope covariance}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept-slope covariance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{intercept-slope covariance, three response categories}} = .002$; $\overline{RC}_{\text{intercept-slope covariance, five response categories}} = .002$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{intercept-slope covariance}}$ in the baseline conditions. Two conditions with threshold non-invariance at two occasions had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept-slope covariance}}$ that did not overlap with that in the corresponding baseline condition, suggesting that these $\overline{RC}_{\text{intercept-slope covariance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{intercept-slope covariance}}$. The difference in $\overline{RC}_{\text{intercept-slope covariance}}$ exceeded .10 between the baseline condition and several conditions with two non-invariant occasions. Thus, a between-subjects ANOVA was conducted on $RC_{\text{intercept-slope covariance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Occasions had a substantial main effect ($\eta^2 = .353$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition with one non-invariant occasion

($\overline{RC}_{\text{intercept-slope covariance, one non-invariant occasion}} = -.014$) than with two non-invariant occasions ($\overline{RC}_{\text{intercept-slope covariance, two non-invariant occasions}} = -.157$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .096$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition with small non-invariance ($\overline{RC}_{\text{intercept-slope covariance, small non-invariance}} = -.048$) than with large non-invariance ($\overline{RC}_{\text{intercept-slope covariance, large non-invariance}} = -.123$). The Number of Non-Invariant Indicators showed a main effect ($\eta^2 = .044$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{RC}_{\text{intercept-slope covariance, one non-invariant indicator}} = -.060$) than when there were three non-invariant indicators ($\overline{RC}_{\text{intercept-slope covariance, three non-invariant indicators}} = -.111$). The Number of Response Categories in the indicators also showed a main effect ($\eta^2 = .042$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition when the indicators had five response categories ($\overline{RC}_{\text{intercept-slope covariance, five response categories}} = -.061$) than when the indicators had three response categories ($\overline{RC}_{\text{intercept-slope covariance, three response categories}} = -.110$).

These main effects were modified by two two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .069$) and a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .052$). When there was one non-invariant occasion, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators were

both negligible. In contrast, when there were two non-invariant occasions, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{\text{intercept-slope covariance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions with three response categories.

Relative changes in the standard errors of the second-order growth parameters. Again, because the mean intercept was constrained to 0 for model identification and the corresponding standard error was 0, I report the relative changes in the standard errors of four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance. I report the relative changes in the standard errors of these growth parameters for completeness. For those second-order growth parameters for which the mean difference in the relative changes exceeded .10 (indicating material bias) between the baseline condition and at least one of the conditions with threshold non-invariance, standard errors are clearly of only secondary interest, and thus the corresponding ANOVA results were not reported.

***Standard error of the mean linear slope*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order mean linear slope ($\overline{RC}_{SE_{\text{mean slope}}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 22 shows the $\overline{RC}_{SE_{\text{mean slope}}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure

represent the corresponding $\overline{RC}_{SE_{\text{mean slope}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{mean slope}},\text{ three response categories}} = .027$; $\overline{RC}_{SE_{\text{mean slope}},\text{ five response categories}} = -.290$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{mean slope}}}$ in the baseline conditions. One condition with large threshold non-invariance for three indicators at two occasions with three response categories had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{mean slope}}}$ that did not overlap with that in the corresponding baseline condition (lower-left panel of Figure 22), suggesting that this $\overline{RC}_{SE_{\text{mean slope}}}$ value differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{mean slope}}}$. The difference in $\overline{RC}_{SE_{\text{mean slope}}}$ exceeded .10 between the baseline condition and one condition with large threshold non-invariance for three indicators at two occasions with three response categories (lower-left panel of Figure 22). When there were three response categories in the indicators as shown in the left panels of Figure 22, in general there was a multiplicative effect of the design factors. When there was one non-invariant indicator, the differences in $\overline{RC}_{SE_{\text{mean slope}}}$ from the corresponding baseline values were all less than .10 (the standard for a meaningful difference), although a small multiplicative effect of the other factors could be observed. When there were three non-invariant indicators, the multiplicative effect of the other design factors became much larger: $\overline{RC}_{SE_{\text{mean slope}}}$ increased when the Magnitude of Non-Invariance was large rather than small and there were two non-invariant occasions rather than one. The $\overline{RC}_{SE_{\text{mean slope}}}$ value

was farthest away from the value in the corresponding baseline condition (.027) when there was large threshold non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the mean linear slope in the conditions with three response categories in the indicators.

When there were five response categories in the indicators as shown in the right panels of Figure 22, the influence of the Number of Non-Invariant Occasions, the Magnitude of Non-Invariance and the Number of Non-Invariant Indicators all appeared to be negligible. This pattern of results was different from the results for the relative changes in the mean linear slope in the conditions with five response categories in the indicators. Although the difference in $\overline{RC}_{\text{mean slope}}$ from the corresponding baseline condition did not exceed the .10 standard in all cases in which there were five response categories, a small multiplicative effect of the other factors on $\overline{RC}_{\text{mean slope}}$ could be observed, such that $\overline{RC}_{\text{mean slope}}$ increased when the Magnitude of Non-Invariance was large rather than small and there were two non-invariant occasions rather than one.

***Standard error of the intercept variance*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept variance ($\overline{RC}_{SE_{\text{intercept variance}}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 23 shows the $\overline{RC}_{SE_{\text{intercept variance}}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept variance}}}$ values in the baseline conditions with fully

invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three response categories}} = -.019$;

$\overline{RC}_{SE_{\text{intercept variance}},\text{ five response categories}} = -.077$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept variance}}}$ in the baseline conditions. Four conditions with three non-invariant indicators had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept variance}}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 23), suggesting that these $\overline{RC}_{SE_{\text{intercept variance}}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{intercept variance}}}$. The difference in $\overline{RC}_{SE_{\text{intercept variance}}}$ exceeded .10 between the baseline condition and two conditions with large threshold non-invariance for three indicators at two occasions (lower panels of Figure 23). Note that the .10 difference standard for a meaningful difference was *not* met for the relative changes in the corresponding growth parameter, the intercept variance, and a between-subjects ANOVA was *not* conducted on the relative changes in the intercept variance. Thus, the ANOVA results on the relative changes in the corresponding *standard error* ($RC_{SE_{\text{intercept variance}}}$) are reported here. Given that the Number of Response Categories had an impact on $RC_{SE_{\text{intercept variance}}}$ in the baseline conditions with fully invariant indicators, a separate between-subjects ANOVA was conducted on $RC_{SE_{\text{intercept variance}}}$ for the threshold non-invariance conditions with three response categories and those with five response categories, respectively.

For conditions with threshold non-invariance and indicators with three response categories (left panels in Figure 23), the ANOVA results showed that the Number of

Non-Invariant Indicators had a substantial main effect ($\eta^2 = .317$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was smaller and closer to the corresponding value in the baseline condition (.019) when there was one non-invariant indicator ($\overline{RC}_{SE_{\text{intercept variance}},\text{ one non-invariant indicator}} = .014$) than when there were three non-invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three non-invariant indicators}} = .056$). The Number of Non-Invariant Occasions showed a main effect ($\eta^2 = .239$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was smaller and closer to the corresponding value in the baseline condition with one non-invariant occasion ($\overline{RC}_{SE_{\text{intercept variance}},\text{ one non-invariant occasion}} = .017$) than with two non-invariant occasions ($\overline{RC}_{SE_{\text{intercept variance}},\text{ two non-invariant occasions}} = .053$). The Magnitude of Non-Invariance also showed a main effect ($\eta^2 = .210$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was smaller and closer to the corresponding value in the baseline condition with small non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\text{ small non-invariance}} = .018$), but became larger and farther away from the corresponding value in the baseline condition with large non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\text{ large non-invariance}} = .052$).

These main effects were modified by three two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .032$), a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .029$), and a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .020$). As shown in the left panels of Figure 23, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant

Occasions were both smaller when there was one non-invariant indicator than when there were three non-invariant indicators. The $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions. Based on the figures, this pattern of results paralleled the results for the relative changes in the intercept variance in the conditions with three response categories in the indicators.

For conditions with threshold non-invariance and indicators with five response categories (right panels in Figure 23), the ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .219$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was more negative and closer to the corresponding value in the baseline condition (-.077) when there was one non-invariant indicator ($\overline{RC}_{SE_{\text{intercept variance}},\text{ one non-invariant indicator}} = -.048$) than when there were three non-invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three non-invariant indicators}} = -.005$). The Number of Non-Invariant Occasions showed a main effect ($\eta^2 = .169$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was more negative and closer to the corresponding value in the baseline condition with one non-invariant occasion ($\overline{RC}_{SE_{\text{intercept variance}},\text{ one non-invariant occasion}} = -.046$) than with two non-invariant occasions ($\overline{RC}_{SE_{\text{intercept variance}},\text{ two non-invariant occasions}} = -.008$). The Magnitude of Non-Invariance also showed a main effect ($\eta^2 = .145$): On average $\overline{RC}_{SE_{\text{intercept variance}}}$ was more negative and closer to the corresponding value in the baseline condition with small non-invariance

($\overline{RC}_{SE_{\text{intercept variance}},\text{ small non-invariance}} = -.044$) than with large non-invariance ($\overline{RC}_{SE_{\text{intercept variance}},\text{ large non-invariance}} = -.009$).

These main effects were modified by three two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .035$), a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .026$), and a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .021$). As shown in the right panels of Figure 23, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both smaller when there was one non-invariant indicator than when there were three non-invariant indicators. The $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions. Based on the figures, this pattern of results paralleled the results for the relative changes in the intercept variance in the conditions with five response categories in the indicators.

Standard error of the linear slope variance. I calculated the relative change (*RC*) in the estimated standard error of the second-order linear slope variance ($\overline{RC}_{SE_{\text{slope variance}}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 24 shows the $\overline{RC}_{SE_{\text{slope variance}}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{slope variance}}}$ values in the baseline conditions with fully

invariant indicators ($\overline{RC}_{SE_{\text{slope variance}},\text{ three response categories}} = -.001$; $\overline{RC}_{SE_{\text{slope variance}},\text{ five response categories}} = -.217$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{slope variance}}}$ in the baseline conditions. One condition with large threshold non-invariance for three indicators at two occasions with three response categories had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{slope variance}}}$ that did not overlap with that in the corresponding baseline condition (lower-left panel of Figure 24), suggesting that this $\overline{RC}_{SE_{\text{slope variance}}}$ value differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{slope variance}}}$. The difference in $\overline{RC}_{SE_{\text{slope variance}}}$ exceeded .10 between the baseline condition and one condition with large threshold non-invariance for three indicators at two occasions with three response categories (lower-left panel of Figure 24). When there were three response categories in the indicators as shown in the left panels of Figure 24, with one non-invariant occasion, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators were both negligible. In contrast, with two non-invariant occasions, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{SE_{\text{slope variance}}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two occasions. This pattern of results paralleled the results for the relative changes in the linear slope variance in the conditions with three response categories in the indicators.

When there were five response categories in the indicators as shown in the right panels of Figure 24, although none of the differences in $\overline{RC}_{SE_{\text{slope variance}}}$ from the corresponding baseline condition met the .10 standard for a meaningful difference, a similar pattern of results was observed. The influence of the other design factors was again negligible with one non-invariant occasion. In contrast, with two non-invariant occasions, the influence of the Magnitude of Non-Invariance became larger, such that the $\overline{RC}_{SE_{\text{slope variance}}}$ value was farther away from the value in the corresponding baseline condition when there was large threshold non-invariance rather than small. This pattern of results paralleled the results for the relative changes in the linear slope variance in the conditions with five response categories in the indicators.

***Standard error of the intercept-slope covariance*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept-slope covariance ($\overline{RC}_{SE_{\text{intercept-slope covariance}}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 25 shows the $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{intercept-slope covariance, three response categories}}} = .009$; $\overline{RC}_{SE_{\text{intercept-slope covariance, five response categories}}} = -.102$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ in the baseline conditions. One condition with large threshold non-

invariance for three indicators at two occasions with three response categories had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ that did not overlap with that in the corresponding baseline condition (lower-left panel of Figure 25), suggesting that this $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$. The difference in $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ exceeded .10 between the baseline condition and one condition with large threshold non-invariance for three indicators with three response categories at two occasions (lower-left panel of Figure 25). When there were three response categories in the indicators as shown in the left panels of Figure 25, with one non-invariant occasion, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators were both negligible. In contrast, with two non-invariant occasions, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the intercept-slope covariance in the conditions with three response categories in the indicators.

When there were five response categories as shown in the right panels of Figure 25, although none of the differences in $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ from the corresponding baseline condition met the .10 standard for a meaningful difference, a similar pattern of results

was observed. With one non-invariant occasion, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators were both negligible. In contrast, with two non-invariant occasions, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the intercept-slope covariance in the conditions with five response categories in the indicators.

Standardized change in the second-order mean linear slope. I calculated the standardized change (*STDC*) in the estimated second-order mean linear slope ($\overline{STDC}_{\text{mean slope}}$) in the conditions with threshold non-invariance, comparing the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. Figure 26 shows the $\overline{STDC}_{\text{mean slope}}$ value with 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{STDC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{STDC}_{\text{mean slope, three response categories}} = .000$; $\overline{STDC}_{\text{mean slope, five response categories}} = -.001$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{STDC}_{\text{mean slope}}$ in the baseline conditions. Several conditions with threshold non-invariance had a 95% normal-theory confidence interval of $\overline{STDC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline

condition, suggesting that these $\overline{STDC}_{\text{mean slope}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{STDC}_{\text{mean slope}}$. The difference in $\overline{STDC}_{\text{mean slope}}$ exceeded .10 between the baseline condition and one condition with large threshold non-invariance for three indicators at two occasions with three response categories (lower-left panel of Figure 26). Thus, a between-subjects ANOVA was conducted on $STDC_{\text{mean slope}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Occasions had a main effect ($\eta^2 = .166$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero and to the value in the corresponding baseline condition with one non-invariant occasion ($\overline{STDC}_{\text{mean slope, one non-invariant occasion}} = -.024$) than with two non-invariant occasions ($\overline{STDC}_{\text{mean slope, two non-invariant occasions}} = -.061$). The Number of Non-Invariant Indicators showed a main effect ($\eta^2 = .159$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero and to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{STDC}_{\text{mean slope, one non-invariant indicator}} = -.024$) than when there were three non-invariant indicators ($\overline{STDC}_{\text{mean slope, three non-invariant indicators}} = -.061$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .113$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero and to the value in the corresponding baseline condition with small non-invariance ($\overline{STDC}_{\text{mean slope, small non-invariance}} = -.027$) than with large non-invariance

($\overline{STDC}_{\text{mean slope, large non-invariance}} = -.058$). The Number of Response Categories in the indicators also showed a main effect ($\eta^2 = .108$): On average $\overline{STDC}_{\text{mean slope}}$ was closer to zero and to the value in the corresponding baseline condition when the indicators had five response categories ($\overline{STDC}_{\text{mean slope, five response categories}} = -.027$) than when the indicators had three response categories ($\overline{STDC}_{\text{mean slope, three response categories}} = -.058$).

These main effects were modified by four two-way interactions: A Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .033$), a Magnitude of Non-Invariance by Number of Non-Invariant Occasions interaction ($\eta^2 = .023$), a Number of Non-Invariant Indicators by Magnitude of Non-Invariance interaction ($\eta^2 = .022$), and a Number of Non-Invariant Indicators by Number of Response Categories interaction ($\eta^2 = .021$). As shown in Figure 26, in general, there was a multiplicative effect of the design factors. When there was one non-invariant indicator, the differences in $\overline{STDC}_{\text{mean slope}}$ from the corresponding baseline values were all less than .10 (the standard for a meaningful difference), although a small multiplicative effect of the other factors could be observed. When there were three non-invariant indicators, the multiplicative effect of the other design factor became much larger: $\overline{STDC}_{\text{mean slope}}$ increased when the Magnitude of Non-Invariance was large rather than small and there were two non-invariant occasions rather than one. In contrast to the earlier results in the loading non-invariance conditions, the use of five rather than three response categories decreased the amount of $\overline{STDC}_{\text{mean slope}}$. The $\overline{STDC}_{\text{mean slope}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large

threshold non-invariance for three indicators at two measurement occasions with three response categories. This pattern of results paralleled the results for the relative changes in the mean linear slope.

Statistical power of the nested model test to detect the incorrect threshold invariance constraints. In the baseline conditions, I computed the empirical Type 1 error rate, i.e. the proportion of the 1000 replications for which a significant test statistic was found, for the nested model test (i.e., DIFFTEST in M*plus*) comparing the model fit of (a) the second-order latent growth model correctly assuming loading invariance, and (b) the second-order latent growth model *correctly* assuming threshold invariance. The empirical Type 1 error rate was .050 when the indicators had three response categories, and .055 when the indicators had five response categories. Since these values were both within the acceptable range of [.0365, .0635], I concluded that this nested model test was not biased in terms of the Type 1 error rate. I then calculated the statistical power of this nested model test in the conditions with manipulated threshold non-invariance, examining the difference in model fit between (a) the second-order latent growth model correctly assuming loading invariance, and (b) the second-order latent growth model *incorrectly* assuming threshold invariance.

Figure 27 shows the statistical power of the nested model test for each condition with threshold non-invariance. As can be seen on the top-right panel of Figure 27, the statistical power to detect *threshold* non-invariance was always above .75. When there was one indicator with small non-invariance in the thresholds and five response categories, the statistical power of the nested model test to detect the incorrect threshold invariance constraints increased as the Number of Non-Invariant Occasions increased. In

all other conditions with threshold non-invariance, the nested model test had very high statistical power (between .989 and 1.00) to detect the incorrect threshold invariance constraints across different combinations of the Number of Non-Invariant Indicators, the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories in the indicators.

Summary. In the conditions with manipulated threshold non-invariance, the Number of Non-Invariant Occasions showed a substantial influence that was greater than that of other design factors on several of the evaluation criteria: The relative changes in the linear slope variance and in the corresponding standard error, and the relative changes in the intercept-slope covariance and the corresponding standard error. With one non-invariant occasion, the influence of the other design factors on these evaluation criteria was negligible. With two non-invariant occasions, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Indicators became much larger: The average evaluation criterion value was farther away from the value in the corresponding baseline condition when there were three non-invariant indicators rather than three, and when the Magnitude of Non-Invariance was large rather than small.

On other evaluation criteria including the relative changes in the mean linear slope and in the corresponding standard error, and the relative change in the standard error of the intercept variance, the Number of Non-Invariant Occasions showed a main effect but it was not substantially greater in magnitude than the influence of other design factors. In general, there was a multiplicative effect of the design factors. When there was one non-invariant indicator, the differences in the mean evaluation criteria from the corresponding baseline values were all less than .10 (the standard for a meaningful

difference), although a small multiplicative effect of the other factors could be observed. When there were three non-invariant indicators, the multiplicative effect of the other design factor became much larger. The differences in the mean evaluation criteria from the corresponding baseline values increased when the Magnitude of Non-Invariance was large rather than small and when there were two non-invariant occasions rather than one. Interestingly, in all cases in which there was either (a) one non-invariant indicator or (b) one non-invariant occasion, the differences in the mean evaluation criteria from the corresponding baseline values did not exceed the .10 standard. The mean evaluation criteria values were farthest away from the values in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two measurement occasions.

It is also noteworthy that in the conditions with threshold non-invariance, the use of five rather than three response categories *decreased* the differences in the evaluation criteria from the corresponding baseline condition. This was contrary to the earlier results in the loading non-invariance conditions, in which the use of five rather than three response categories *increased* the differences in the evaluation criteria from the corresponding baseline condition. Note that there were two thresholds in an indicator with three response categories, and four thresholds in an indicator with five response categories. Threshold non-invariance was introduced on the last *one* threshold of the non-invariant indicator(s). Thus, a plausible explanation of the greater effect of threshold non-invariance when there were three response categories is that the proportion of problematic thresholds in a non-invariant indicator was 50% when there were three response categories, but only 25% when there were five response categories.

In addition, when the indicators had five response categories, the 95% normal-theory confidence intervals of the components of evaluation criteria 1) and 2) were always much wider than when the indicators had three response categories, and this was true for both the conditions with non-invariant thresholds and the baseline conditions with fully invariant indicators. This result parallels the findings in the conditions with loading non-invariance, and may be related to the larger number of parameters to estimate in the configural and loading invariance models when there were five response categories in the indicators rather than three.

Also, when the indicators had five response categories, the standard errors of the growth parameters always decreased substantially as a result of adding correct threshold invariance constraints in the baseline conditions with fully invariant indicators. When the added threshold invariance constraints were incorrect (i.e., in the threshold non-invariance conditions), almost all of the standard errors of the growth parameters tended to decrease, and these relative change values did not differ substantially from the corresponding values in the baseline conditions. There was one exception: On average the standard error of the intercept variance *increased* as a result of adding incorrect threshold invariance constraints in the condition with large threshold non-invariance on three indicators at two occasions with fiver response categories. One implication is that the conclusion of statistical significance of a growth parameter of interest may change as more invariance constraints were added, whether or not the added invariance constraints were appropriate. In contrast, when the indicators had three response categories, the standard errors of the growth parameters did not change substantially as a result of adding correct threshold invariance constraints in the baseline conditions. These standard errors

sometimes changed substantially as a result of adding *incorrect* threshold invariance constraints in the threshold non-invariance conditions: The change tended to be positive for the standard error of the intercept variance, but negative for the standard errors of the other growth parameters for the present population model.

The nested model test comparing the fit of the second-order latent growth model assuming loading invariance with that of the model assuming threshold invariance had acceptable Type 1 error rates. Contrary to the earlier results where the nested model test had very low statistical power (less than .15) to detect *loading* non-invariance for one indicator, the statistical power to detect *threshold* non-invariance was always above .75. When there was one indicator with small thresholds non-invariance and five response categories, the statistical power of this nested model test to detect the incorrect threshold invariance constraints increased as the Number of Non-Invariant Occasions increased. In all other conditions with threshold non-invariance, the nested model test had very high statistical power (above .95) to detect the incorrect threshold invariance constraints.

Conditions with Unique Factor Non-Invariance

Relative changes in the second-order growth parameters. Again, since the mean intercept was constrained to 0 for model identification, I report the relative changes in four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance.

***Mean linear slope*.** I calculated the relative change (*RC*) in the estimated second-order mean linear slope ($\overline{RC}_{\text{mean slope}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 28 shows the $\overline{RC}_{\text{mean slope}}$ value with 95%

normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{mean slope, three response categories}} = .003$; $\overline{RC}_{\text{mean slope, five response categories}} = .001$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{mean slope}}$ in the baseline conditions. None of the unique factor non-invariance conditions had a 95% normal-theory confidence interval of $\overline{RC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was *not* met for $\overline{RC}_{\text{mean slope}}$. The greatest difference in $\overline{RC}_{\text{mean slope}}$ between the baseline condition and a condition with manipulated unique factor non-invariance was .074. Thus, a between-subjects ANOVA was *not* conducted on $RC_{\text{mean slope}}$.

Intercept variance. I calculated the relative change (*RC*) in the estimated second-order intercept variance ($\overline{RC}_{\text{intercept variance}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 29 shows the $\overline{RC}_{\text{intercept variance}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{intercept variance, three response categories}} = .004$; $\overline{RC}_{\text{intercept variance, five response categories}} = -.001$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for

$\overline{RC}_{\text{intercept variance}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept variance}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was met for $\overline{RC}_{\text{intercept variance}}$. The difference in $\overline{RC}_{\text{intercept variance}}$ exceeded .10 between the baseline condition and several conditions with three non-invariant indicators (lower panels of Figure 29). Thus, a between-subjects ANOVA was conducted on $RC_{\text{intercept variance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators showed a main effect ($\eta^2 = .186$): On average $\overline{RC}_{\text{intercept variance}}$ was smaller and closer to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{RC}_{\text{intercept variance, one non-invariant indicator}} = .046$) than when there were three non-invariant indicators ($\overline{RC}_{\text{intercept variance, three non-invariant indicators}} = .135$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .086$): On average $\overline{RC}_{\text{intercept variance}}$ was smaller and closer to the value in the corresponding baseline condition with small non-invariance ($\overline{RC}_{\text{intercept variance, small non-invariance}} = .060$) than with large non-invariance ($\overline{RC}_{\text{intercept variance, large non-invariance}} = .121$). The Number of Non-Invariant Occasions also showed a main effect ($\eta^2 = .061$): On average $\overline{RC}_{\text{intercept variance}}$ was smaller and closer to the value in the corresponding baseline condition with one non-invariant occasion

($\overline{RC}_{\text{intercept variance, one non-invariant occasion}} = .065$) than with two non-invariant occasions ($\overline{RC}_{\text{intercept variance, two non-invariant occasions}} = .116$).

These main effects were modified by a two-way interaction between the Number of Non-Invariant Indicators and the Number of Non-Invariant Occasions interaction ($\eta^2 = .022$). As shown in Figure 29, when there was one non-invariant indicator, the influence of the Number of Non-Invariant Occasions was rather small. In contrast, when there were three non-invariant indicators, the influence of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{\text{intercept variance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large unique factor non-invariance for three indicators at two measurement occasions.

Linear slope variance. I calculated the relative change (*RC*) in the estimated second-order linear slope variance ($\overline{RC}_{\text{slope variance}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 30 shows the $\overline{RC}_{\text{slope variance}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{slope variance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{slope variance, three response categories}} = .011$; $\overline{RC}_{\text{slope variance, five response categories}} = .004$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{slope variance}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had

a 95% normal-theory confidence interval of $\overline{RC}_{\text{slope variance}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was met for $\overline{RC}_{\text{slope variance}}$. The difference in $\overline{RC}_{\text{slope variance}}$ exceeded .10 between the baseline condition and several conditions with three non-invariant indicators (lower panels of Figure 30). Thus, a between-subjects ANOVA was conducted on $RC_{\text{slope variance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Occasions had a main effect ($\eta^2 = .164$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition with one non-invariant occasion ($\overline{RC}_{\text{slope variance, one non-invariant occasion}} = -.028$) than with two non-invariant occasions ($\overline{RC}_{\text{slope variance, two non-invariant occasions}} = -.119$). The Number of Non-Invariant Indicators showed a main effect ($\eta^2 = .136$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{RC}_{\text{slope variance, one non-invariant indicator}} = -.032$) than when there were three non-invariant indicators ($\overline{RC}_{\text{slope variance, three non-invariant indicators}} = -.115$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .027$): On average $\overline{RC}_{\text{slope variance}}$ was closer to zero and to the value in the corresponding baseline condition with small non-invariance ($\overline{RC}_{\text{slope variance, small non-invariance}} = -.055$) than with large non-invariance ($\overline{RC}_{\text{slope variance, large non-invariance}} = -.092$).

These main effects were modified by a two-way interaction between the Number of Non-Invariant Indicators and the Number of Non-Invariant Occasions ($\eta^2 = .038$). As shown in Figure 30, when there was one non-invariant occasion, the influence of the Number of Non-Invariant Indicators was very small. In contrast, when there were two non-invariant occasions, the influence the Number of Non-Invariant Indicators became much larger. The $\overline{RC}_{\text{slope variance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large threshold non-invariance for three indicators at two occasions with three response categories.

Intercept-slope covariance. I calculated the relative change (*RC*) in the estimated second-order intercept-slope variance ($\overline{RC}_{\text{intercept-slope covariance}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 31 shows the $\overline{RC}_{\text{intercept-slope covariance}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{\text{intercept-slope covariance}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{\text{intercept-slope covariance, three response categories}} = .011$; $\overline{RC}_{\text{intercept-slope covariance, five response categories}} = .006$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{\text{intercept-slope covariance}}$ in the baseline conditions. Two conditions with unique factor non-invariance for three indicators had a 95% normal-theory confidence interval of $\overline{RC}_{\text{intercept-slope covariance}}$ that did not overlap with that in the corresponding baseline condition (lower panels of Figure 31),

suggesting that these $\overline{RC}_{\text{intercept-slope covariance}}$ values differed from the corresponding baseline value at a level of significance that is at least $\alpha = .01$.

The .10 difference standard for a meaningful difference was also met for $\overline{RC}_{\text{intercept-slope covariance}}$. The difference in $\overline{RC}_{\text{intercept-slope covariance}}$ exceeded .10 between the baseline condition and all conditions with three non-invariant indicators and several conditions with one non-invariant indicator. Thus, a between-subjects ANOVA was conducted on $RC_{\text{intercept-slope covariance}}$ to provide information about the importance of each of the factors in the design.

The ANOVA results showed that the Number of Non-Invariant Indicators had a substantial main effect ($\eta^2 = .418$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition when there was one non-invariant indicator ($\overline{RC}_{\text{intercept-slope covariance, one non-invariant indicator}} = -.088$) than when there were three non-invariant indicators ($\overline{RC}_{\text{intercept-slope covariance, three non-invariant indicators}} = -.302$). The Magnitude of Non-Invariance showed a main effect ($\eta^2 = .138$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition with small non-invariance ($\overline{RC}_{\text{intercept-slope covariance, small non-invariance}} = -.133$) than with large non-invariance ($\overline{RC}_{\text{intercept-slope covariance, large non-invariance}} = -.256$). The Number of Non-Invariant Occasions showed a main effect ($\eta^2 = .067$): On average $\overline{RC}_{\text{intercept-slope covariance}}$ was closer to zero and to the value in the corresponding baseline condition with one non-invariant occasion

($\overline{RC}_{\text{intercept-slope covariance, one non-invariant occasion}} = -.152$) than with two non-invariant occasions ($\overline{RC}_{\text{intercept-slope covariance, two non-invariant occasions}} = -.238$).

These main effects were modified by two two-way interactions: A Magnitude of Non-Invariance by Number of Non-Invariant Indicators interaction ($\eta^2 = .041$) and a Number of Non-Invariant Indicators by Number of Non-Invariant Occasions interaction ($\eta^2 = .023$). When there was one non-invariant indicator, the influence of the Magnitude of Non-Invariance and the influence of the Number of Non-Invariant Occasions were both negligible. In contrast, when there were three non-invariant indicators, the influence of the Magnitude of Non-Invariance and that of the Number of Non-Invariant Occasions became much larger. The $\overline{RC}_{\text{intercept-slope covariance}}$ value was farthest away from zero and from the value in the corresponding baseline condition when there was large unique factor non-invariance for three indicators at two measurement occasions.

Relative changes in the standard errors of the second-order growth parameters. Again, because the mean intercept was constrained to 0 for model identification and the corresponding standard error was 0, I report the relative changes in the standard errors of four second-order latent growth parameters below: mean linear slope, intercept variance, linear slope variance, and intercept-slope covariance. I report the relative changes in the standard errors of these growth parameters for completeness. For those second-order growth parameters for which the mean difference in the relative changes exceeded .10 (indicating material bias) between the baseline condition and at least one of the conditions with unique factor non-invariance, standard errors are clearly of only secondary interest, and thus the corresponding ANOVA results were not reported.

***Standard error of the mean linear slope*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order mean linear slope ($\overline{RC}_{SE_{\text{mean slope}}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 32 shows the $\overline{RC}_{SE_{\text{mean slope}}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{mean slope}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{mean slope}},\text{ three response categories}} = -.384$; $\overline{RC}_{SE_{\text{mean slope}},\text{ five response categories}} = -.368$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{mean slope}}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{mean slope}}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was *not* met for $\overline{RC}_{SE_{\text{mean slope}}}$. The greatest difference in $\overline{RC}_{SE_{\text{mean slope}}}$ between the baseline condition and a condition with unique factor non-invariance was .018. Thus, a between-subjects ANOVA was *not* conducted on $RC_{SE_{\text{mean slope}}}$.

***Standard error of the intercept variance*.** I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept variance ($\overline{RC}_{SE_{\text{intercept variance}}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure

33 shows the $\overline{RC}_{SE_{\text{intercept variance}}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept variance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{intercept variance}},\text{ three response categories}} = -.287$; $\overline{RC}_{SE_{\text{intercept variance}},\text{ five response categories}} = -.253$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept variance}}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept variance}}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was met for $\overline{RC}_{SE_{\text{intercept variance}}}$. The difference in $\overline{RC}_{SE_{\text{intercept variance}}}$ exceeded .10 between the baseline condition and two conditions with large unique factor non-invariance for three indicators at two occasions (lower panels of Figure 33). When there were three response categories in the indicators as shown in the left panels of Figure 33, the difference in $\overline{RC}_{SE_{\text{intercept variance}}}$ between a condition with unique factor non-invariance and the corresponding baseline condition was greater when there were three non-invariant indicators rather than one, when the Magnitude of Non-Invariance was large rather than small, and when there were two non-invariant occasions rather than one. The $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large unique factor non-invariance for three indicators at two measurement occasions. When there were five response categories in the indicators as shown in the right panels of Figure 33, the

difference in $\overline{RC}_{SE_{\text{intercept variance}}}$ between a condition with unique factor non-invariance and the corresponding baseline condition was greater when there were three non-invariant indicators rather than one, when the Magnitude of Non-Invariance was large rather than small, and when there were two non-invariant occasions rather than one. The influence of the Number of Non-Invariant Occasions was greater when there were three non-invariant indicators rather than one. Again the $\overline{RC}_{SE_{\text{intercept variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was large unique factor non-invariance for three indicators at two measurement occasions. This pattern of results paralleled the results for the relative changes in the intercept variance.

Standard error of the linear slope variance. I calculated the relative change (*RC*) in the estimated standard error of the second-order linear slope variance ($\overline{RC}_{SE_{\text{slope variance}}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 34 shows the $\overline{RC}_{SE_{\text{slope variance}}}$ value with the 95% normal-theory confidence limits for each condition with unique factor non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{slope variance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{slope variance}},\text{ three response categories}} = -.190$; $\overline{RC}_{SE_{\text{slope variance}},\text{ five response categories}} = -.176$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{slope variance}}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence

interval of $\overline{RC}_{SE_{\text{slope variance}}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was met for $\overline{RC}_{SE_{\text{slope variance}}}$. The difference in $\overline{RC}_{SE_{\text{slope variance}}}$ exceeded .10 between the baseline condition and one condition with large unique factor non-invariance for three indicators at two occasions with three response categories (lower-left panel of Figure 34). When there were three response categories in the indicators as shown in the left panels of Figure 34, the difference in $\overline{RC}_{SE_{\text{slope variance}}}$ between a condition with unique factor non-invariance and the corresponding baseline condition was greater when there were three non-invariant indicators rather than one and when there were two non-invariant occasions rather than one. The $\overline{RC}_{SE_{\text{slope variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was unique factor non-invariance for three indicators at two occasions. When there were five response categories in the indicators as shown in the right panels of Figure 34, although none of the differences in $\overline{RC}_{SE_{\text{slope variance}}}$ from the corresponding baseline condition met the .10 standard for a meaningful difference, a similar pattern of results was observed. The difference in $\overline{RC}_{SE_{\text{slope variance}}}$ between a condition with unique factor non-invariance and the corresponding baseline condition was greater when there were three non-invariant indicators rather than one and when there were two non-invariant occasions rather than one. The $\overline{RC}_{SE_{\text{slope variance}}}$ value was farthest away from the value in the corresponding baseline condition when there was unique factor non-invariance for three indicators at two occasions. This pattern of results

differed from the results for the relative changes in the linear slope variance, in that the Magnitude of Non-Invariance had an influence on the relative changes in the linear slope variance, but not on the relative changes in the standard error of the linear slope variance.

Standard error of the intercept-slope covariance. I calculated the relative change (*RC*) in the estimated standard error of the second-order intercept-slope covariance ($\overline{RC}_{SE_{\text{intercept-slope covariance}}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 35 shows the $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ value with the 95% normal-theory confidence limits for each condition with threshold non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ values in the baseline conditions with fully invariant indicators ($\overline{RC}_{SE_{\text{intercept-slope covariance}},\text{ three response categories}} = -.093$; $\overline{RC}_{SE_{\text{intercept-slope covariance}},\text{ five response categories}} = -.094$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence interval of $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was *not* met for $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$. The greatest difference in $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$ between the baseline condition and a condition with unique factor non-invariance was .036. Thus, a between-subjects ANOVA was *not* conducted on $\overline{RC}_{SE_{\text{intercept-slope covariance}}}$.

Standardized change in the second-order mean linear slope. I calculated the standardized change (*STDC*) in the estimated second-order mean linear slope ($\overline{STDC}_{\text{mean slope}}$) in the conditions with unique factor non-invariance, comparing the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. Figure 36 shows the $\overline{STDC}_{\text{mean slope}}$ value with 95% normal-theory confidence limits for each condition with loading non-invariance. As a benchmark, the solid lines in the figure represent the corresponding $\overline{STDC}_{\text{mean slope}}$ values in the baseline conditions with fully invariant indicators ($\overline{STDC}_{\text{mean slope, three response categories}} = .002$; $\overline{STDC}_{\text{mean slope, five response categories}} = .000$), and the dotted lines represent the corresponding 95% normal-theory confidence limits for $\overline{STDC}_{\text{mean slope}}$ in the baseline conditions. None of the conditions with unique factor non-invariance had a 95% normal-theory confidence interval of $\overline{STDC}_{\text{mean slope}}$ that did not overlap with that in the corresponding baseline condition.

The .10 difference standard for a meaningful difference was *not* met for $\overline{STDC}_{\text{mean slope}}$. The greatest difference in $\overline{STDC}_{\text{mean slope}}$ between the baseline condition and a condition with unique factor non-invariance was .062. Thus, a between-subjects ANOVA was *not* conducted on $STDC_{\text{mean slope}}$.

Statistical power of the nested model test to detect the incorrect unique factor invariance constraints. In the baseline conditions, I computed the empirical Type 1 error rates, i.e. the proportions of the 1000 replications for which a significant test statistic was found, for the nested model test comparing the model fit of (a) the second-order latent

growth model correctly assuming threshold invariance, and (b) the second-order latent growth model *correctly* assuming unique factor invariance. The empirical Type 1 error rate was .057 when the indicators had three response categories, and .061 when the indicators had five response categories. Since these values were both within the acceptable range of [.0365, .0635], I concluded that this nested model test was not biased in terms of the Type 1 error rate. I then calculated the statistical power of this nested model test in the conditions with unique factor non-invariance, examining the difference in model fit between (a) the second-order latent growth model correctly assuming threshold invariance, and (b) the second-order latent growth model *incorrectly* assuming unique factor invariance. Figure 37 shows the statistical power of the nested model test for each condition with unique factor non-invariance. As can be seen on the figure, the statistical power to detect *unique factor* non-invariance was always above .45. When there was one indicator with small non-invariance in the unique factor variance (upper panels of Figure 37), the statistical power of the nested model test to detect the incorrect unique factor invariance constraints increased when there were two non-invariant occasions rather than one. The statistical power was also slightly higher when there were five response categories in the indicators rather than three. In all other conditions with unique factor non-invariance, the nested model test had very high statistical power (between .992 and 1.00) to detect the incorrect unique factor invariance constraints across different combinations of the Number of Non-Invariant Indicators, the Magnitude of Non-Invariance, the Number of Non-Invariant Occasions, and the Number of Response Categories in the indicators.

Summary. In the conditions with unique factor non-invariance, imposing incorrect longitudinal unique factor invariance constraints did not have much impact on the relative changes in the mean linear slope or the standardized changes in the mean linear slope. On the other hand, imposing incorrect longitudinal unique factor invariance constraints tended to lead to positive relative changes in the intercept variance, negative relative changes in the linear slope variance, and negative relative changes in the intercept-slope covariance. In general, there was a multiplicative effect of the design factors. The influence of the Number of Non-Invariant Occasions was smaller when there was one non-invariant indicator than when there were three non-invariant indicators. For the relative change in the intercept-slope covariance only, the influence of the Magnitude of Non-Invariance was also smaller when there was one non-invariant indicator rather than three. The differences from the corresponding baseline values in the mean relative change in the growth parameters increased when the Magnitude of Non-Invariance was large rather than small and when there were two non-invariant occasions rather than one.

Contrary to the earlier results in the loading non-invariance conditions and the threshold non-invariance conditions, the influence of the Number of Response Categories in the indicators was trivial in the conditions with unique factor non-invariance: The differences between the average evaluation criterion values and the corresponding baseline value were similar whether the indicators had five response categories or three. Regarding the standard errors of the growth parameters, they always decreased substantially as a result of adding unique factor invariance constraints, regardless of the number of response categories, or whether the added unique factor invariance constraints were correct or incorrect. A plausible explanation of the lack of influence of the Number

of Response Categories in the conditions with unique factor non-invariance is that the two fitted second-order latent growth models assumed either threshold invariance or unique factor invariance at the first-order level. Therefore, the parameters directly influenced by the number of response categories, i.e., the threshold parameters, were always constrained to be equal over time in the conditions with unique factor non-invariance.

The nested model test comparing the fit of the second-order latent growth model assuming threshold invariance with that of the model assuming unique factor invariance had acceptable Type 1 error rates. The statistical power to detect *unique factor* non-invariance was always above .45. When there was one indicator with a small magnitude of non-invariance in the unique factor variance, the statistical power of the nested model test to detect the incorrect unique factor invariance constraints increased when there were two non-invariant occasions rather than one. The statistical power was also slightly higher when there were five response categories in the indicators rather than three. In all other conditions with unique factor non-invariance, the nested model test had very high statistical power (above .95) to detect the incorrect unique factor invariance constraints.

CHAPTER 9

DISCUSSION

This dissertation investigated the effect of longitudinal measurement non-invariance on parameter estimates and standard errors in second-order latent growth models. The focus was on models with ordered-categorical indicators. In the second-order latent growth models used in this dissertation study, the first-order component is the measurement model comprised of ordered-categorical indicators. The second-order component is a linear growth model of the latent common factors of the continuous normally distributed latent responses assumed to underlie the observed ordered-categorical indicators. The goal of this dissertation was to examine the suitability of using the second-order latent growth model to gauge the practical importance of longitudinal measurement non-invariance. Put differently, the research question was whether growth parameters in the second-order latent growth model would be seriously biased if the researcher acts as if the indicators achieve longitudinal measurement invariance, when in fact longitudinal measurement invariance is not achieved. If the estimated values of the growth parameters in the second-order latent growth model change following the addition of incorrect longitudinal measurement invariance constraints, then such changes can be viewed as a sensitivity analysis measure of longitudinal measurement non-invariance.

Numerous forms of sensitivity analysis measures have been proposed to evaluate the practical importance of the measurement non-invariance of binary or ordered categorical data across groups, and may be generalized to provide information on the practical importance of the measurement non-invariance over time. To provide some examples, under the item response theory framework, for instance, Stark, Chernyshenko,

and Drasgow (2004) proposed sensitivity analyses relating measurement non-invariance of the whole scale to mean raw score differences across groups or to selection ratios and cut scores for selection decisions. Steinberg and Thissen (2006) proposed calculating the standardized difference in the metric of item parameters. Meade (2010) proposed several sensitivity analysis measures derived from the expected score of an indicator or from the expected scale (or test) score. Under the structural equation modeling framework, Nye and Drasgow (2011) proposed an indicator-level sensitivity measure of the standardized difference between groups, and derived equations for the influence of measurement non-invariance on the mean and variance of the whole scale. Kuha and Moustaki (2015) examined the distortions in the estimated means and variances of the latent common factors across multiple groups for binary data, as a result of loading or threshold non-invariance. Oberski (2014) proposed a sensitivity analysis for multigroup structural equation models of ordinal data, calculating the expected changes in the structural parameters if inappropriate measurement invariance constraints were to be freed. Such studies of sensitivity analyses of measurement non-invariance, as well as studies of statistical tests to detect measurement non-invariance, have typically focused on the influence of non-invariant factor loadings and non-invariant threshold parameters, but have assumed that unique factor variances were always invariant (e.g., Gonzalez-Roma et al., 2006; Kim & Yoon, 2011; Kuha & Moustaki, 2015; Stark et al., 2006). In popular structural equation modeling programs, by default the unique factor variances (or the total variances) of the ordinal indicators were typically constrained to be equal to a unit value (and thus invariant across groups or across time). However, Liu et al. (in press) showed that unique factor invariance is a necessary condition to attribute mean changes

over time in the observed ordered-categorical indicators entirely to changes in the latent construct. Moreover, in longitudinal studies the unique factor variances are likely to vary across time. From evidence with continuous data, constraining unique factor variances to be equal across time when in fact they vary can result in bias in the estimated latent variable covariance parameters like the variances of the latent intercept and latent slope and their covariance in a latent linear growth model (Kwok, West, & Green, 2007).

This dissertation separately simulated longitudinal measurement non-invariance in three different locations in the confirmatory factor model: factor loadings, thresholds, and unique factor variances. In addition, I examined factors that can potentially affect the magnitude of the influence of measurement non-invariance, including the number of non-invariant indicators (e.g., Kuha & Moustaki, 2015; Meade & Lautenschlager, 2004) and the magnitude of non-invariance per indicator (e.g., Gonzalez-Roma et al., 2006; Kuha & Moustaki, 2015; Stark et al., 2006). Unique to the investigation of measurement invariance in longitudinal designs, I examined the influence of the number of non-invariant occasions. I also examined the influence of the number of response categories per indicator, which has been shown to influence common factor models of ordinal data (e.g., Rhemtulla et al., 2012) and which has been examined in some previous simulation studies of measurement invariance in binary or ordered categorical data across groups (e.g., Kim & Yoon, 2011; Stark et al., 2006). The examination of the influence of unique factor non-invariance and the number of non-invariant occasions, and the interaction of these factors with the other design factors, should provide new insights into the practical importance of violations of longitudinal measurement invariance.

For each simulated data set with measurement non-invariance, two second-order latent growth models were fitted. The first model assumed the correct level of longitudinal measurement invariance (the baseline model). The second model assumed an incorrect level of longitudinal measurement invariance that was one level higher in the hierarchy. The convergence rates were generally very high. The worst convergence rates (91.7%-93.5%) were found for the second-order latent growth models assuming configural invariance (which had the greatest number of parameters to be estimated) in the conditions with large loading non-invariance, where there was a combination of low factor loadings and relatively sparse response categories at the last measurement occasion.

In this dissertation simulation, I examined the relative changes in the growth parameters and in their corresponding standard errors between these two models. I also examined the standardized change in the mean linear slope, calculated as the magnitude of change in the mean linear slope relative to the square root of the intercept variance in the correctly specified baseline model. The intercept variance in the correctly specified baseline model represents an estimate of the population variance at the reference occasion. The standardized change in the mean linear slope may be more informative than the relative change in the mean linear slope when the mean linear slope is close to zero.

Sensitivity of the Different Growth Parameters

This dissertation found that each growth parameter in the second-order latent growth model was *differentially* sensitive to the location of non-invariance. The relative change in the mean linear slope and the standardized change in the mean linear slope

were sensitive to longitudinal non-invariance in the factor loadings and in the thresholds, but *not* sensitive to the simulated longitudinal non-invariance in the unique factor variances. The relative change in the intercept variance, on the other hand, was sensitive to longitudinal non-invariance in the unique factor variances, but *not* sensitive to longitudinal non-invariance in the factor loadings or in the thresholds. The relative changes in the slope variance and in the intercept-slope covariance were sensitive to longitudinal non-invariance in the factor loadings, in the thresholds, and in the unique factor non-invariance. When a specific growth parameter was sensitive to a certain location of non-invariance, the magnitude of the relative change or standardized change in the growth parameter depended on a multiplicative function of the Number of Non-Invariant Indicators, the Number of Non-Invariant Occasions, the Magnitude of Non-Invariance, and the Number of Response Categories in the indicators.

Most Prominent Design Factor in Different Locations of Non-Invariance

Given a particular location of non-invariance, the pattern of the influence of the other design factors was consistent for those growth parameters sensitive to this location of non-invariance. Across different locations of non-invariance, however, the design factor with the most prominent influence was different.

Loading Non-Invariance. When longitudinal measurement non-invariance only occurred in the factor loadings, the Number of Non-Invariant Indicators had the most prominent influence. With one non-invariant indicator, the relative change or standardized change in those growth parameters sensitive to loading non-invariance (i.e., mean linear slope, slope variance, intercept-slope covariance) was always negligible. With three non-invariant indicators, the magnitude of the relative change or standardized

change became much larger, and the influence of the Number of Non-Invariant Occasions, the Magnitude of Non-Invariance, and the Number of Response Categories became more evident. The magnitude of the relative change or standardized change was greater when there were two non-invariant occasions rather than one, when the magnitude of non-invariance was large rather than small, and when the indicators had five response categories rather than three.

Threshold Non-Invariance. When longitudinal measurement non-invariance only occurred in the thresholds, the Number of Non-Invariant Occasions had the most prominent influence. With one non-invariant occasion, the relative change or standardized change in a growth parameter sensitive to threshold non-invariance (i.e., mean linear slope, slope variance, intercept-slope covariance) was always negligible. With two non-invariant occasions, the magnitude of the relative change or standardized change became much larger, and the influence of the other design factors became more evident. The magnitude of the relative change or standardized change was greater when there were three non-invariant indicators rather than one, when the magnitude of non-invariance was large rather than small, and when the indicators had three response categories rather than five.

Unique Factor Non-Invariance. When the longitudinal measurement non-invariance only occurred in the unique factor variances, the Number of Non-Invariant Indicators had a substantial influence. With three non-invariant indicators rather than one, the relative change in those growth parameters sensitive to unique factor non-invariance (i.e., intercept variance, slope variance, intercept-slope covariance) was always greater, and the multiplicative effect of the other design factors became more evident. The

magnitude of the relative change was greater when there were two non-invariant occasions rather than one and when the magnitude of non-invariance was large rather than small. The Number of Response Categories did not have any influence when the longitudinal measurement non-invariance only occurred in the unique factor variances.

Influence of the Number of Response Categories on the Growth Parameter Estimates

The results for the design factor Number of Response Categories in the indicators were particularly noteworthy because the influence of this design factor was strikingly different across different locations of non-invariance. When longitudinal measurement non-invariance occurred only in the factor loadings, the magnitude of the relative change or standardized change in the growth parameters on average tended to be greater when the indicators had five response categories rather than three. When longitudinal measurement non-invariance occurred only in the thresholds, the magnitude of the relative change or standardized change in the growth parameters on average tended to be *smaller* when the indicators had five response categories rather than three. When longitudinal measurement non-invariance occurred only in the unique factor variances, the magnitude of the relative change or standardized change in the growth parameters on average was similar whether the indicators had five response categories or three. Given that threshold non-invariance was introduced on the last threshold of the non-invariant indicator(s) in the simulation, a plausible explanation of the greater effect of threshold non-invariance when there were three response categories is that the proportion of problematic thresholds in a non-invariant indicator was 50% when there were three response categories with two thresholds, but only 25% when there were five response

categories with four thresholds. In the conditions with unique factor non-invariance, two second-order latent growth models were fitted, one correctly assuming threshold invariance and the other incorrectly assuming unique factor invariance. Thus a plausible explanation of the lack of influence of the Number of Response Categories in the conditions with unique factor non-invariance is that the parameters directly influenced by the number of response categories, i.e., the threshold parameters, were always constrained to be equal over time.

The Importance of Unique Factor Invariance

The mathematical derivations in Appendix A proved that unique factor invariance is a necessary condition to attribute mean changes over time in the observed ordered-categorical indicators to changes in the latent construct. However, the magnitudes of longitudinal unique factor non-invariance examined in this dissertation study did not lead to a substantial change in the estimated mean linear slope in the second-order latent growth model. The average relative change in the estimated mean linear slope after adding the incorrect unique factor invariance constraints reached its maximum discrepancy from the corresponding value in the baseline condition (-7.1%) when there was large non-invariance for three indicators at the last two measurement occasions. In this dissertation study, large non-invariance in the unique factor variances was defined as having the unique variances become 2.25 times as large at the non-invariant occasion(s) as compared to the first occasion. It is possible that a larger magnitude of non-invariance in the unique factor variances will have a greater influence on the estimated mean linear slope. To provide a better understanding of the influence of unique factor non-invariance, I conducted some additional simulations with a very large sample size (N = 1,000,000)

using the same population parameter values for data generation as in Table 2. These simulations provide a large-sample comparison of the models. When all indicators were fully invariant over time, the ratios of the unique factor variance to the total variance for indicators X_3, X_4, and X_5 ranged from .17 to .26 at the third occasion, and from .12 to .18 at the fourth occasion. When the unique factor variances for indicators X_3, X_4, and X_5 became 2.25 times as large (i.e., the large non-invariance condition in the original dissertation simulations) at the third and fourth measurement occasions as compared to the first occasion[13], the relative change in the estimated mean linear slope was -6.3%, similar to the results in the original dissertation simulations. With a rather extreme magnitude of non-invariance such that the unique factor variances became 9.0 times as large for three of the indicators at the third and fourth measurement occasions as compared to the first occasion[14], the relative change in the mean linear slope was -11.5%. These results suggest that for researchers who are only interested in the mean linear slope, the magnitude of unique factor non-invariance needs to be very large to have a material influence on the estimated mean linear slope.

On the other hand, the influence of longitudinal unique factor non-invariance on the other growth parameters was much larger. In the original dissertation simulations, on average the relative changes in the intercept variance and in the slope variance deviated from zero by more than 20% when the unique factor variances became 2.25 times as large for three of the indicators at the third and fourth measurement occasions as

[13] The corresponding ratios of the unique factor variance to the total variance for these non-invariant indicators ranged from .32 to .44 at the third occasion, and from .23 to .33 at the fourth occasion.

[14] The corresponding ratios of the unique factor variance to the total variance for these non-invariant indicators ranged from .65 to .75 at the third occasion, and from .55 to .67 at the fourth occasion.

compared to the first occasion. On average the relative change in the intercept-slope covariance under these conditions deviated from zero by around 50%. Similar magnitudes of the relative changes were obtained in the additional simulation with a very large (N = 1,000,000) sample when the unique factor variances were 2.25 times as large for three of the indicators at the third and fourth measurement occasions as compared to the first occasion. In the additional simulation in which the unique factor variance was 9.0 times as large at the third and fourth occasions compared to the first occasion for three of the indicators, the relative change in the intercept variance was 147.5%, the relative change in the slope variance was -21.8%, and the relative change in the intercept-slope covariance was -214.5%. These results suggest that researchers interested in explaining the intercept variance, the slope variance, or the intercept-slope covariance would clearly need to take into account longitudinal unique factor non-invariance.

Influence of the Number of Non-Invariant Occasions

This dissertation study examined the influence of the number of non-invariant occasions. This design factor is unique to studies of longitudinal measurement invariance. In general, having two non-invariant occasions rather than one led to greater changes in the growth parameter estimates, especially with a larger number of non-invariant indicators and a larger magnitude of non-invariance per indicator. The influence of the number of non-invariant occasions was most prominent when longitudinal measurement non-invariance occurred in the threshold parameters.

Standard Errors of the Growth Parameters

Researchers studying growth over time are often interested in the statistical significance of the growth parameters of interest. Because the statistical test of a growth

parameter of interest depends on both the parameter estimate and the corresponding standard error, I also examined the influence of inappropriate longitudinal measurement invariance constraints on the standard errors of the growth parameters.

When Each Indictor Had Three Response Categories. In the baseline conditions with fully invariant indicators, the standard errors of the growth parameters changed minimally as a result of adding *correct* loading invariance or threshold invariance constraints, but decreased substantially as a result of adding *correct* unique factor invariance constraints. In the conditions with measurement non-invariance, the relative changes in the standard errors following the addition of *incorrect* invariance constraints tended to be *negatively biased*[15] relative to the corresponding baseline condition for the standard errors of the mean linear slope, the slope variance, and the intercept-slope covariance. On the other hand, the relative change in the standard error of the *intercept variance* following the addition of *incorrect* invariance constraints tended to *positively biased*[16] relative to the corresponding baseline condition.

[15] Here "negatively biased" refers to the fact that when adding, say, *correct* loading invariance constraints in the baseline condition led to *no change* in the standard error of a growth parameter, adding *incorrect* loading invariance constraints in the loading non-invariance constraints tended to lead to a *decrease* in the corresponding standard error. Similarly, when adding *correct* unique factor invariance constraints in the baseline condition led to a *decrease* in the standard error, adding *incorrect* unique factor invariance constraints in the unique factor non-invariance conditions tended to lead to a *greater decrease* in the corresponding standard error.

[16] Here "positively biased" refers to the fact that when adding, say, *correct* loading invariance constraints in the baseline condition led to *no change* in the standard error of the intercept variance, adding *incorrect* loading invariance constraints in the loading non-invariance constraints tended to lead to an *increase* in the corresponding standard error. Similarly, when adding *correct* unique factor invariance constraints in the baseline condition led to a *decrease* in the standard error of the intercept variance, adding *incorrect* unique factor invariance constraints in the unique factor non-invariance conditions tended to lead to a *greater decrease* in the corresponding standard error.

When Each Indictors Had Five Response Categories. In the baseline conditions with fully invariant indicators, the standard errors of the growth parameters always decreased substantially as a result of adding *correct* invariance constraints. In the conditions with loading non-invariance, the relative changes in the standard errors following the addition of *incorrect* loading invariance constraints tended to be *negatively biased* relative to the corresponding baseline condition. In the conditions with threshold or unique factor non-invariance, the relative changes in the standard errors following the addition of *incorrect* invariance constraints tended to be *negatively biased* relative to the corresponding baseline condition for the standard errors of the mean linear slope, the slope variance, and the intercept-slope covariance, but *positively biased* for the standard error of the intercept variance.

One implication is that the conclusion of statistical significance or non-significance of a growth parameter of interest may change as more invariance constraints are added, whether or not the added invariance constraints are appropriate. This finding implies that statistical *significance* or non-significance of a growth parameter of interest should *not* be used as a criterion for assessing the practical importance of longitudinal measurement non-invariance. Instead, researchers should focus on effect size measures representing the *magnitude* of change in the growth parameters as a result of imposing incorrect longitudinal measurement invariance constraints.

The Nested Model Test

In this dissertation study, all the nested model likelihood ratio tests that compared the fit of the two second-order latent growth models with different longitudinal

measurement invariance constraints had an acceptable empirical Type 1 error rate[17] in the baseline conditions with fully invariant indicators. In conditions where there were three non-invariant indicators, the likelihood ratio test had very high statistical power[18] (> .95), regardless of the location of measurement non-invariance. However, in conditions where there was one non-invariant indicator, the statistical power of the likelihood ratio test depended on the location of non-invariance, which is in line with findings in the studies of measurement non-invariance across groups (e.g., Gonzalez-Roma et al., 2006; Kim & Yoon, 2011). These results suggest that the likelihood ration test of the nested models is differentially sensitive to different locations of non-invariance, which further highlights the importance of examining the magnitude of change in the growth parameters as a result of incorrect longitudinal measurement invariance constraints.

[17] The empirical Type 1 error rate is represented by the proportion of the 1000 replications for which a significant nested model test statistic is found in a baseline condition with fully invariant indicators.

[18] The statistical power is represented by the proportion of the 1000 replications for which a significant nested model test statistic is found in a condition with manipulated measurement non-invariance.

CHAPTER 10

LIMITATIONS AND IMPLICATIONS FOR FUTURE RESEARCH

The proposed sensitivity analysis relies heavily on the accuracy of the parameter estimates from the second-order latent growth model with ordered-categorical indicator. Therefore, when this sensitivity analysis is applied to real data sets where the population model is unknown, factors that affect the accuracy of parameter estimates from the second-order latent growth model will also potentially confound interpretations of this sensitivity analysis.

Of importance, the proposed sensitivity analysis makes the assumption that the appropriate specification of the growth model was used in the second-order latent growth models. When applying this sensitivity analysis to real data sets, if an incorrect growth model is specified, then the estimated growth parameters may be biased even with correct longitudinal measurement invariance constraints. This caveat applies both to research using ordered-categorical indicators and to research using continuous indicators. For instance, with continuous indicators, Murphy, Beretvas, and Pituch (2011) investigated how growth parameters in a second-order latent growth model could be influenced by an unmodeled autoregressive or autoregressive and moving average process among the first-ordered latent common factors. They found that the mean intercept and mean slope parameters were unbiased, but the intercept variance, the slope variance, and the intercept-slope covariance tended to be biased, especially with high (.8) or moderate (.5) unmodeled autocorrelation. In addition, Wirth (2008) found that model fit statistics from latent growth models using composites of continuous indicators or factor scores saved from measurement models with inappropriate invariance constraints tended to have

increased chances of accepting an alternative form to the true form of growth. An indirect implication for the sensitivity analysis proposed in this dissertation study is that if the functional form of growth is misspecified in the second-order latent growth models, then changes in the estimated growth parameters following the addition of incorrect measurement invariance constraints may provide a confounded depiction of the practical importance of the violation of measurement invariance.

A second caveat in applying this sensitivity analysis to real data sets is that second-order latent growth models are more likely to produce inadmissible solutions than first-order latent growth models that do not include a measurement model. This caveat applies both to research using ordered-categorical indicators and to research using continuous indicators (Grimm et al, in press, Chapter 15; Leite, 2007). Using a larger sample size may alleviate this problem (Leite, 2007).

A third caveat in applying this sensitivity analysis to real data sets with ordered-categorical indicators is that if the bivariate or multivariate frequency table of the ordered-categorical indicators has sparse or empty cells, estimation of polychoric correlations may be problematic (Brown & Bendetti, 1977; Flora & Curran, 2004; Bollen & Curran, 2006). This problem can potentially influence the parameter estimates from the second-order latent growth model. This problem is especially relevant in longitudinal studies, because with population level mean change over time, sparse data are likely to occur for the lowest or highest response categories at the earliest or latest measurement occasions.

This dissertation simulation study used a large sample size (N = 2000) and specified the correct functional form of growth in the second-order latent growth models.

I chose parameter values in the population data generation model such that the simulated data were not too sparse. The lowest expected cell count in the *univariate* frequency table with five response categories per indicator was around 80, and the lowest expected cell count in the *bivariate* frequency table was around 5 in this dissertation study. Further research is needed to investigate the influence of improper specification of the growth model, the influence of sample size, and the influence of sparse data on this sensitivity analysis.

A fourth caveat is that I simulated measurement non-invariance to occur in only one location at a time. Longitudinal measurement non-invariance was simulated to occur either only in the factor loadings, only in the thresholds, or only in the unique factor variances. This strategy provided a clear picture of the influence of longitudinal measurement non-invariance in different locations. However, in practice, it is possible to have measurement non-invariance occurring simultaneously in more than one location.

CHAPTER 11

CONCLUDING REMARKS

This study examined how sensitive the second-order latent growth parameters are to different locations of longitudinal measurement non-invariance with ordered-categorical indicators, and explored the influence of a number of factors including the Number of Non-Invariant Indicators, the Number of Non-Invariant Occasions, the Magnitude of Non-Invariance, and the Number of Response Categories. The results of this dissertation study suggested that for researchers only interested in describing the average linear growth trajectory, longitudinal loading non-invariance and longitudinal threshold non-invariance can each have a substantial influence on the estimate of the mean growth trajectory. In contrast, longitudinal unique factor non-invariance needs to reach a rather extreme magnitude to have a material influence on the estimate of the mean growth trajectory. For researchers interested in explaining the intercept variance, longitudinal unique factor non-invariance can have a substantial influence, whereas longitudinal loading non-invariance and longitudinal threshold non-invariance have only a minimal influence on the estimate of the intercept variance. For researchers interested in explaining the slope variance or the intercept-slope covariance, longitudinal measurement non-invariance in the factor loadings, in the thresholds, and in the unique factor variances can all influence the corresponding estimated growth parameters. Effects of non-invariance depend on the location of non-invariance, on the number of response categories (for loading non-invariance and threshold non-invariance), and on the various factors determining the total degree of non-invariance in the model. These factors include

the number of non-invariant indicators, the number of non-invariant occasions, and the magnitude of non-invariance for each non-invariant indicator.

Table 1

Population Parameters of the Simulation Study in the Baseline Conditions with Fully Invariant Indicators – Before Transformation

First-order level model	λ	ν		$\sigma^2_{jj(t)}$	$\rho_{jj(t,t+1)}$
Indicator *j*		Three response categories	Five response categories		
X_{1t}	1.00	[-0.05, 1.05]	[-0.05, 0.35, 0.75, 1.05]	0.30	0.20
X_{2t}	0.80	[-0.50, 0.65]	[-0.50, -0.10, 0.25, 0.65]	0.30	0.20
X_{3t}	0.90	[0.05, 1.15]	[0.05, 0.50, 0.85, 1.15]	0.30	0.20
X_{4t}	0.70	[-0.80, 0.40]	[-0.80, -0.40, 0.00, 0.40]	0.30	0.20
X_{5t}	0.80	[-0.55, 0.85]	[-0.55, -0.05, 0.45, 0.85]	0.30	0.20
Second-order level model	$\gamma_{\text{intercept}}$	γ_{slope}	Disturbance variance	$R^2_{\eta_1}$	
First-order latent common factors					
η_1	1	0	0.21	0.70	
η_2	1	1	0.33	0.70	
η_3	1	2	0.54	0.70	
η_4	1	3	0.83	0.70	
Second-order latent growth factors	ξ_0	ξ_1			
Mean	0	0.6			
Variance/ Covariance	0.50				
	0.089	0.10			

Note. At the first-order level, all unique factor correlations of lag-1 were set to $\rho_{jj(t,t+1)} = \rho$, all unique factor correlations of lag-2 were set to $\rho_{jj(t,t+2)} = \rho^2$, and all unique factor correlations of lag-3 were set to $\rho_{jj(t,t+3)} = \rho^3$. All lagged disturbance correlations of the first-order latent common factors were set to zero.

Table 2

Population Parameters of the Simulation Study in the Baseline Conditions with Fully Invariant Indicators – After Transformation

First-order level model	λ	ν		$\sigma^2_{jj(t)}$	$\rho_{jj(t,t+1)}$
Indicator *j*		Three response categories	Five response categories		
X_{1t}	1.00	[-0.09, 1.92]	[-0.09, 0.64, 1.37, 1.92]	1.00	0.20
X_{2t}	0.80	[-0.91, 1.19]	[-0.91, -0.18, 0.46, 1.19]	1.00	0.20
X_{3t}	0.90	[0.09, 2.10]	[0.09, 0.91, 1.55, 2.10]	1.00	0.20
X_{4t}	0.70	[-1.46, 0.73]	[-1.46, -0.73, 0.00, 0.73]	1.00	0.20
X_{5t}	0.80	[-1.00, 1.55]	[-1.00, -0.09, 0.82, 1.55]	1.00	0.20
Second-order level model	$\gamma_{\text{intercept}}$	γ_{slope}	Disturbance variance	$R^2_{\eta_1}$	
First-order latent common factors					
η_1	1	0	0.71	0.70	
η_2	1	1	1.11	0.70	
η_3	1	2	1.79	0.70	
η_4	1	3	2.75	0.70	
Second-order latent growth factors	ξ_0	ξ_1			
Mean	0	1.09			
Variance/ Covariance	1.66				
	0.296	0.33			

Note. At the first-order level, all unique factor correlations of lag-1 were set to $\rho_{jj(t,t+1)} = \rho$, all unique factor correlations of lag-2 were set to $\rho_{jj(t,t+2)} = \rho^2$, and all unique factor correlations of lag-3 were set to $\rho_{jj(t,t+3)} = \rho^3$. All lagged disturbance correlations of the first-order latent common factors were set to zero.

Table 3

Proportions of Non-Convergence of the Configural Invariance Model and Computational Problems for the Baseline Conditions

Number of response categories	% Non-convergence of the configural invariance model	% Computational Problems
3	5.3%	0.70%
5	4.2%	0.10%

Note. The computational problem here refers to the situation in which the DIFFTEST comparing the fit of the model assuming loading invariance versus the model assuming configural invariance could not be computed.

Table 4

Proportions of Non-Convergence of the Configural Invariance Model and Computational Problems for the Conditions with Loading Non-Invariance

Magnitude of non-invariance	Number of non-invariant indicators	Number of occasions with non-invariant indicators	Number of response categories	% Non-convergence of the configural invariance model	% Computational Problems
Small	1	1	3	4.6%	0.30%
Small	1	1	5	4.4%	0.00%
Small	1	2	3	6.2%	0.30%
Small	1	2	5	6.4%	0.30%
Small	3	1	3	5.4%	0.30%
Small	3	1	5	3.5%	0.20%
Small	3	2	3	4.7%	0.70%
Small	3	2	5	4.1%	0.70%
Large	1	1	3	7.0%	0.30%
Large	1	1	5	8.3%	0.40%
Large	1	2	3	7.8%	0.20%
Large	1	2	5	8.1%	0.40%
Large	3	1	3	7.6%	0.70%
Large	3	1	5	7.1%	0.30%
Large	3	2	3	6.5%	0.30%
Large	3	2	5	7.3%	0.20%

Note. The computational problem here refers to the situation in which the DIFFTEST comparing the fit of the model assuming loading invariance versus the model assuming configural invariance could not be computed.

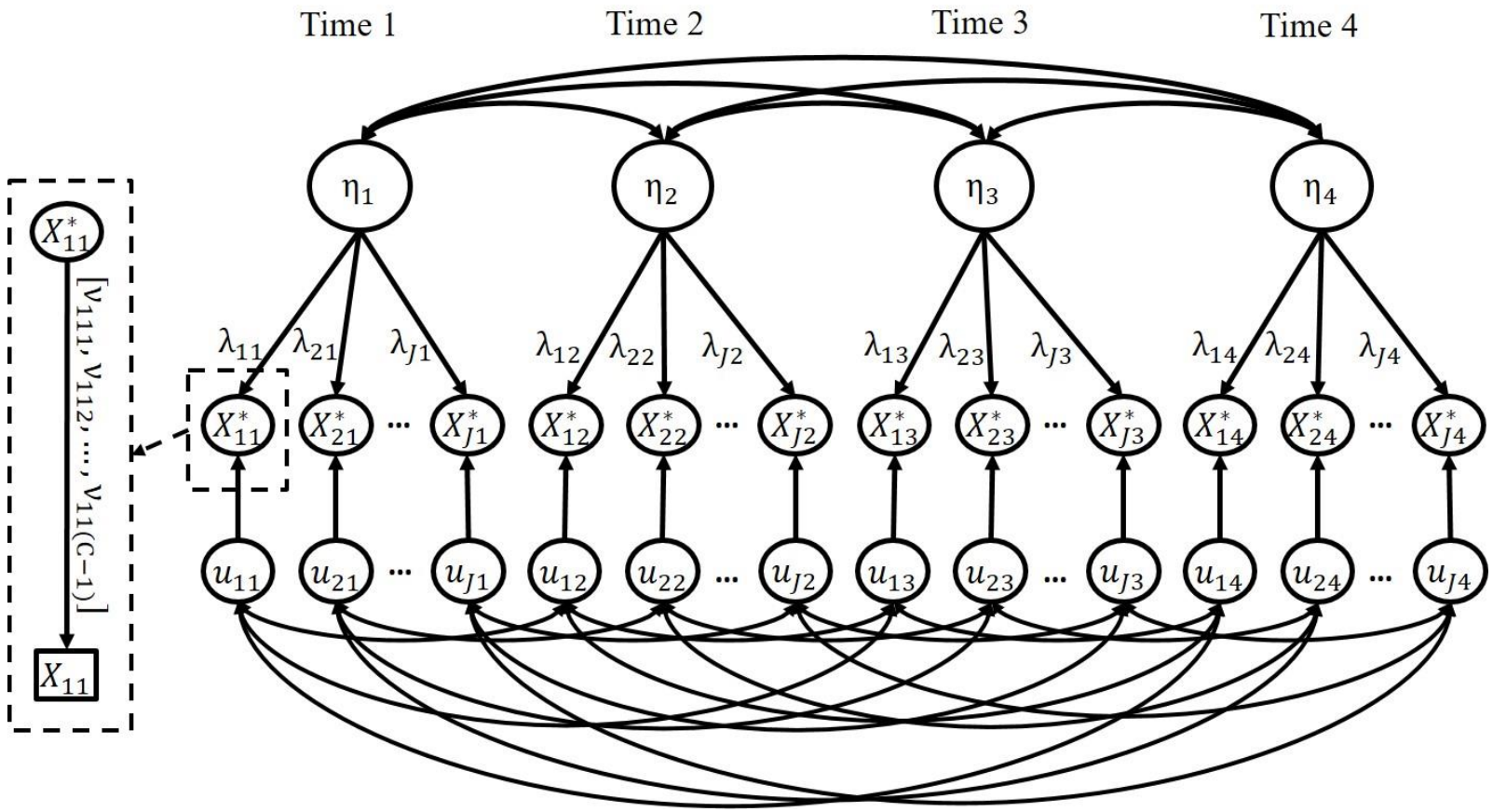

Figure 1. Longitudinal ordered-categorical CFA model.

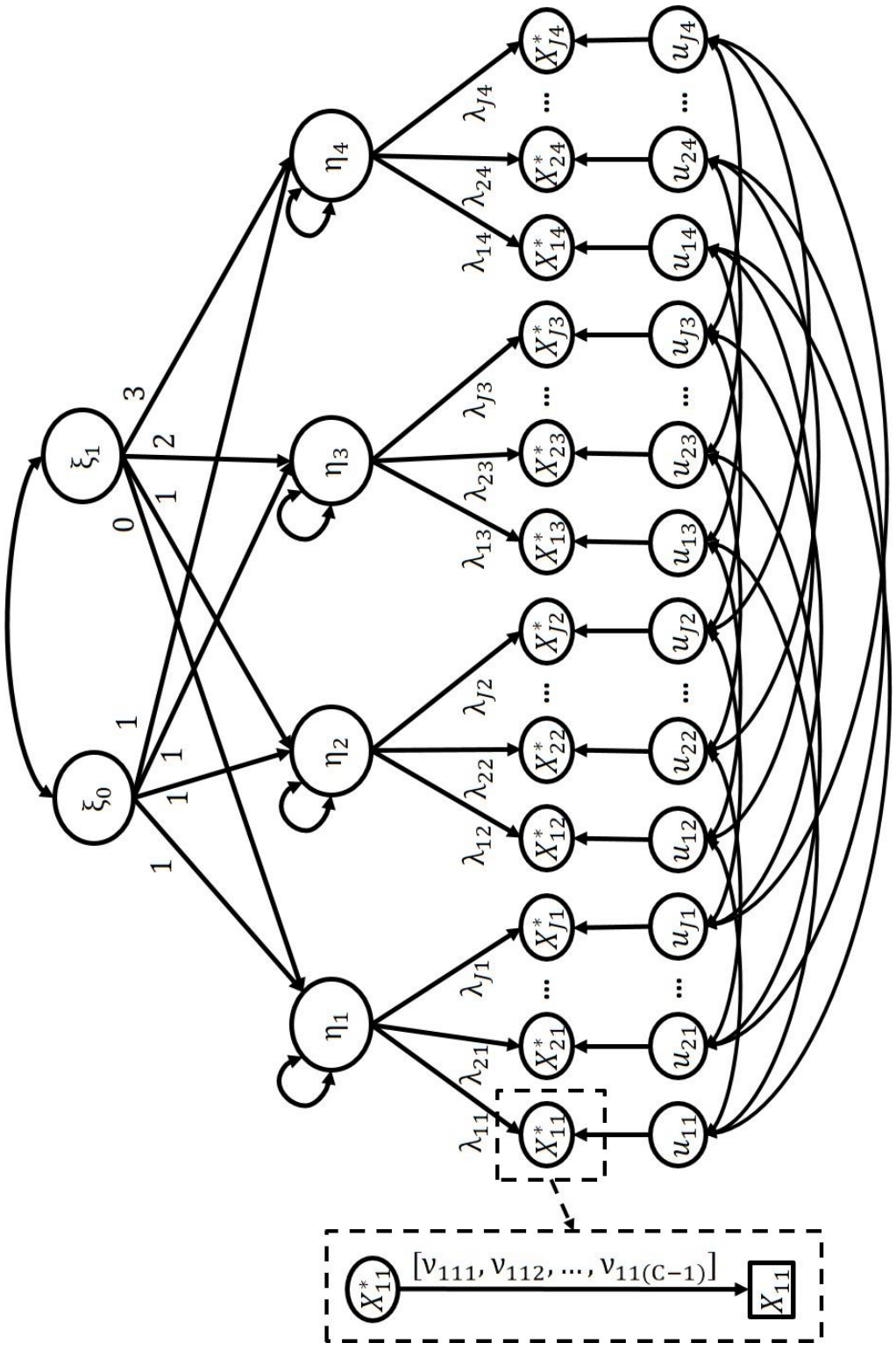

Figure 2. Second-order latent linear growth model with ordinal

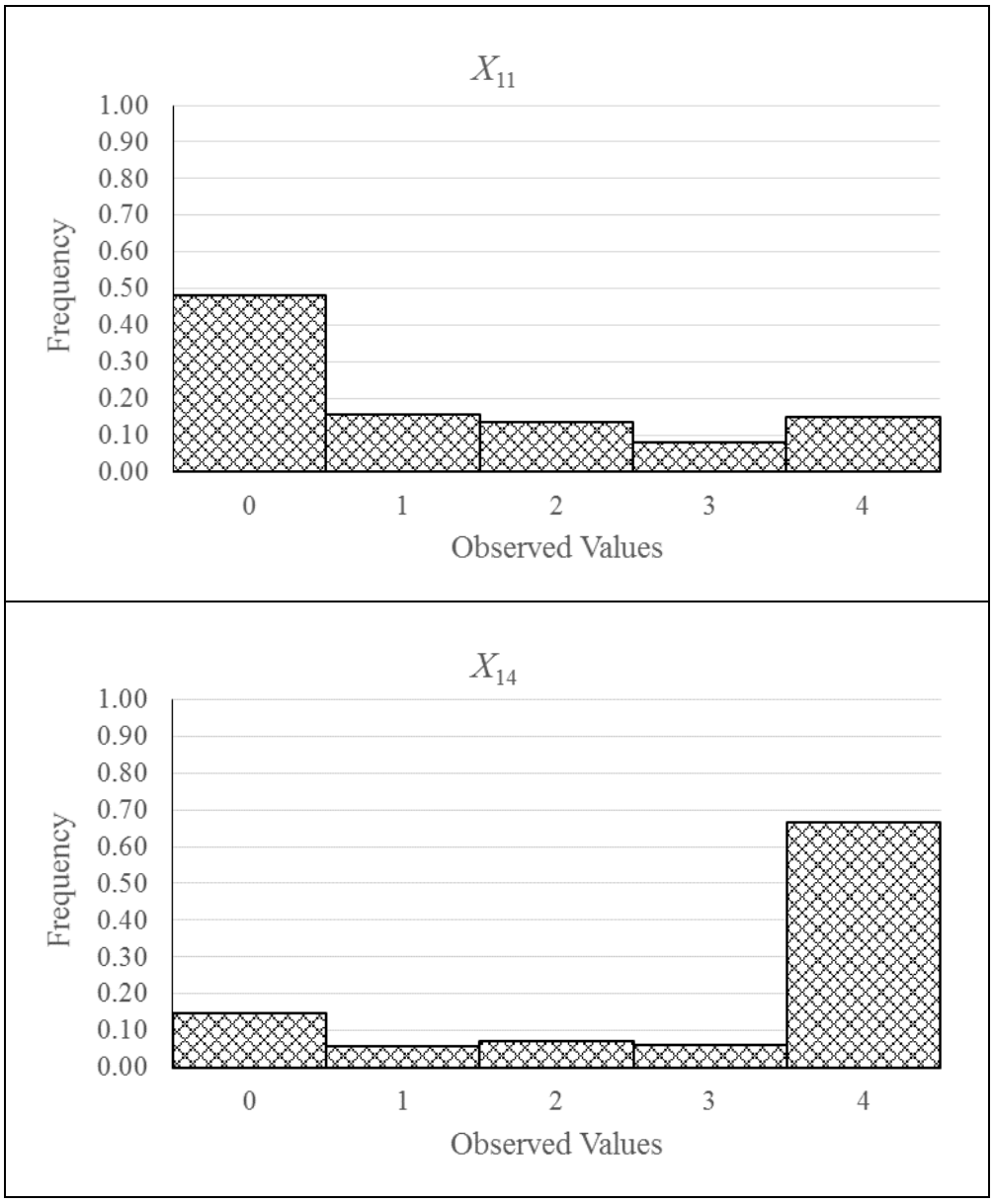

Figure 3. Observed distribution of indicator X_1 in the baseline condition with five response categories. *Note*: The upper panel contains the distribution at the first measurement occasion, and the lower panel contains the distribution at the last measurement occasion.

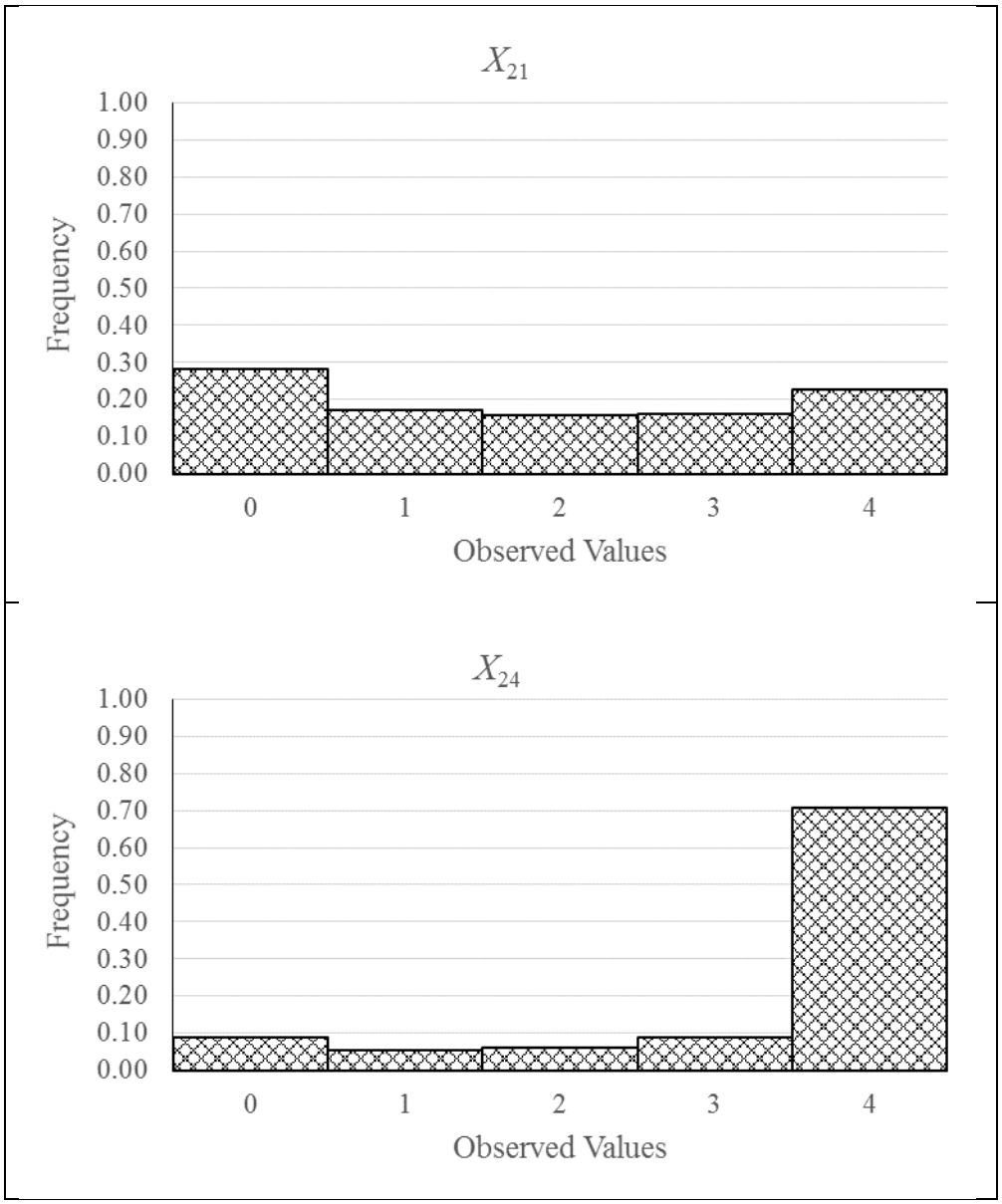

Figure 4. Observed distribution of indicator X_2 in the baseline condition with five response categories. *Note*: The upper panel contains the distribution at the first measurement occasion, and the lower panel contains the distribution at the last measurement occasion.

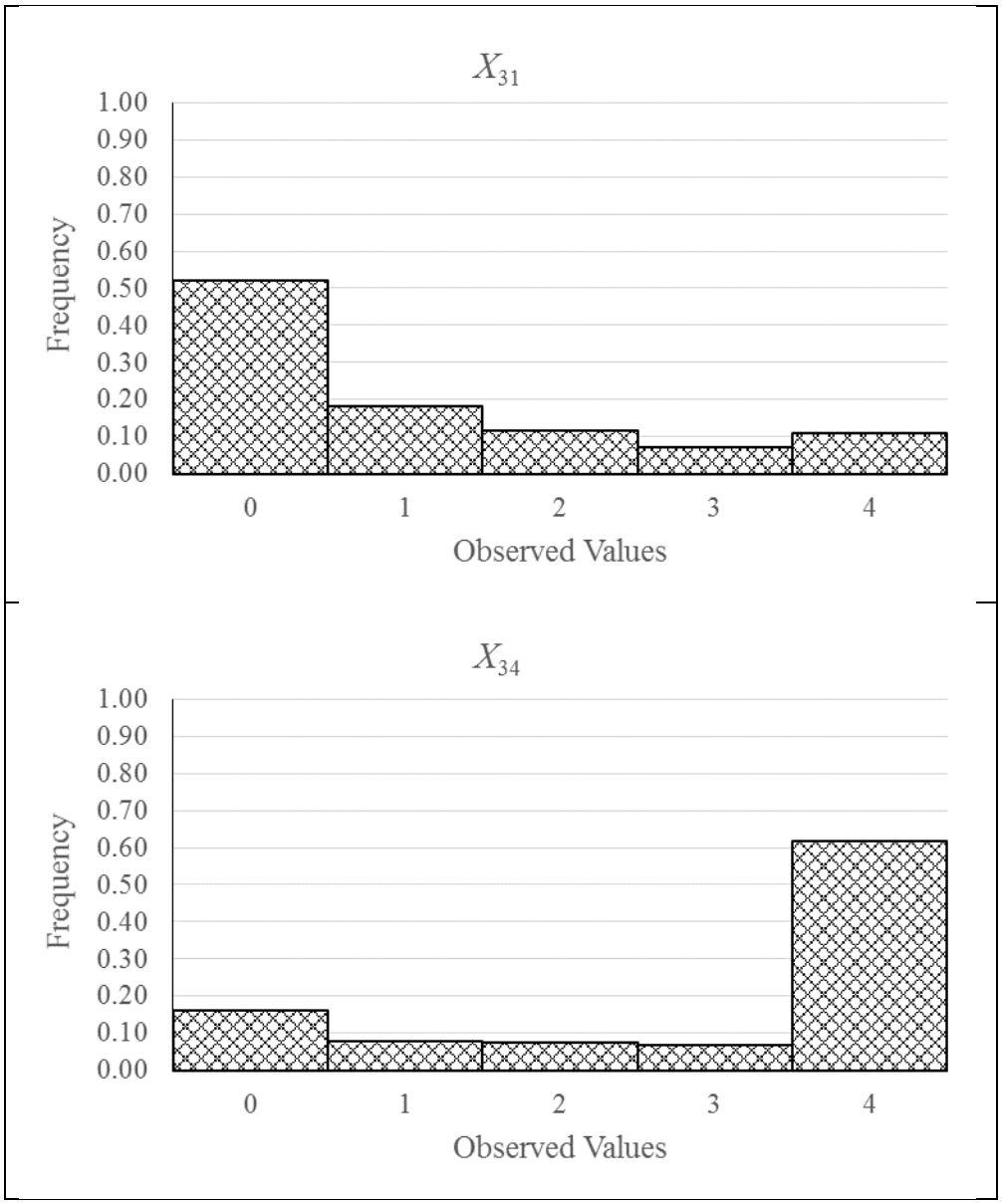

Figure 5. Observed distribution of indicator X_3 in the baseline condition with five response categories. *Note*: The upper panel contains the distribution at the first measurement occasion, and the lower panel contains the distribution at the last measurement occasion.

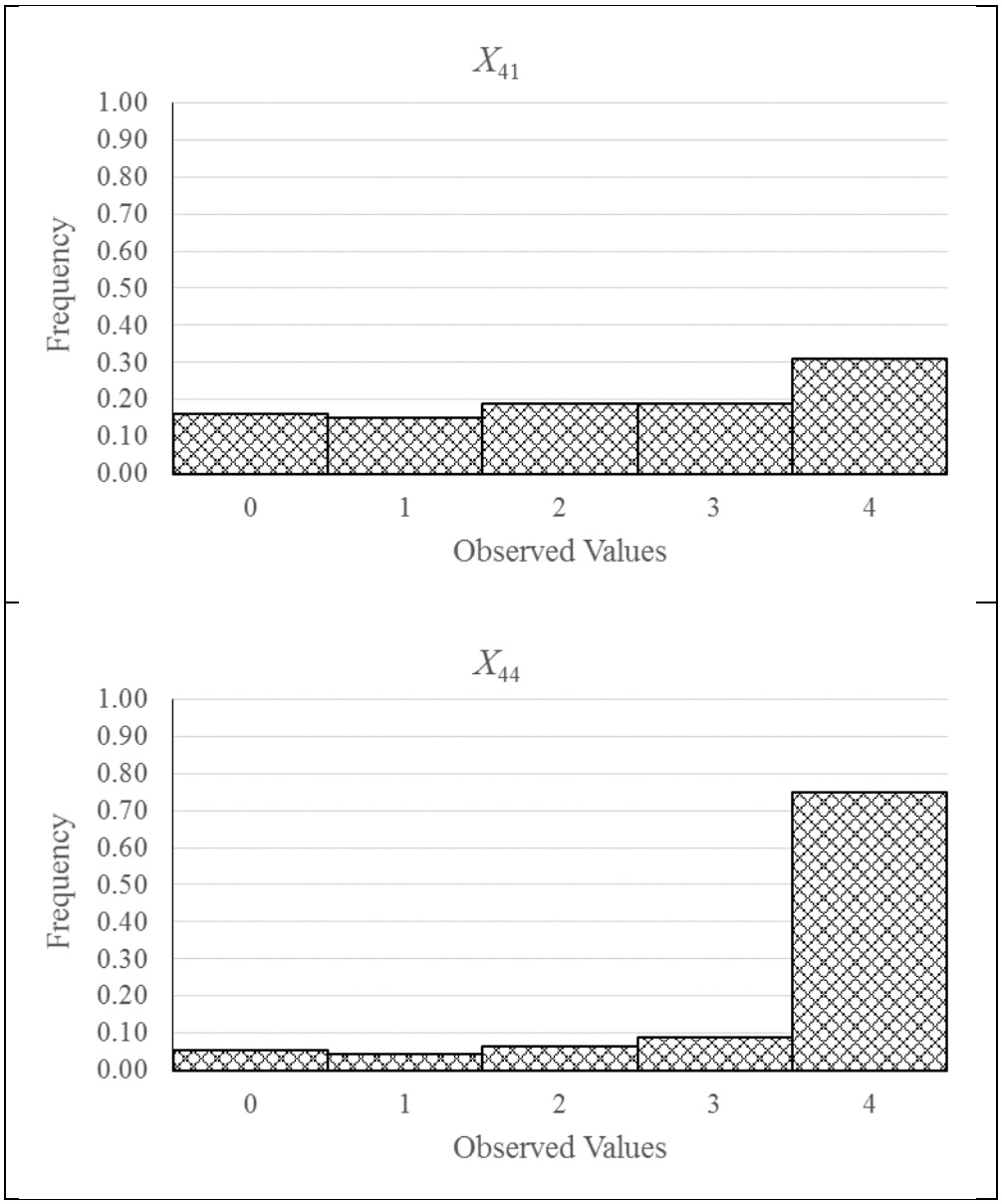

Figure 6. Observed distribution of indicator X_4 in the baseline condition with five response categories. *Note*: The upper panel contains the distribution at the first measurement occasion, and the lower panel contains the distribution at the last measurement occasion.

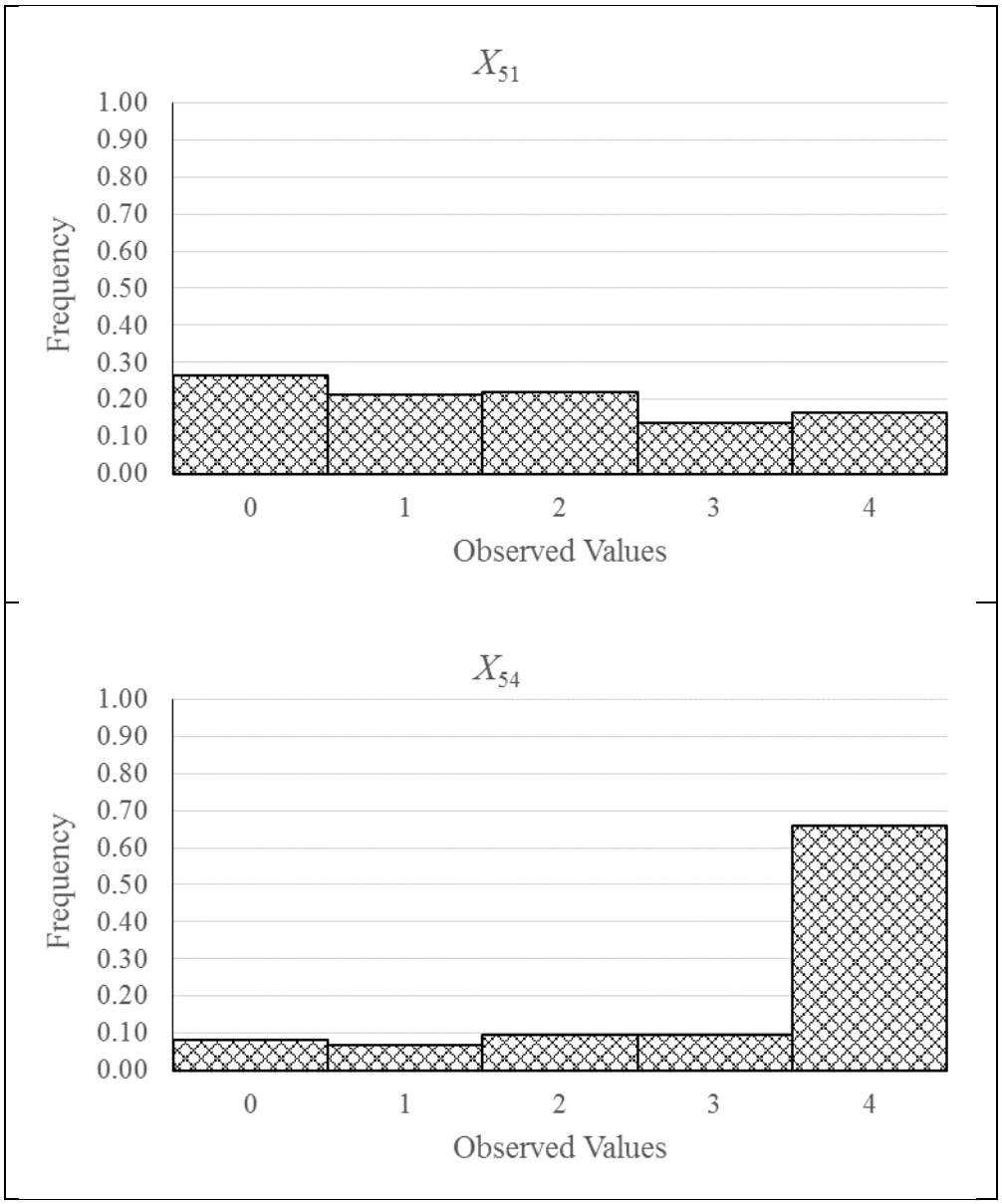

Figure 7. Observed distribution of indicator X_5 in the baseline condition with five response categories. *Note*: The upper panel contains the distribution at the first measurement occasion, and the lower panel contains the distribution at the last measurement occasion.

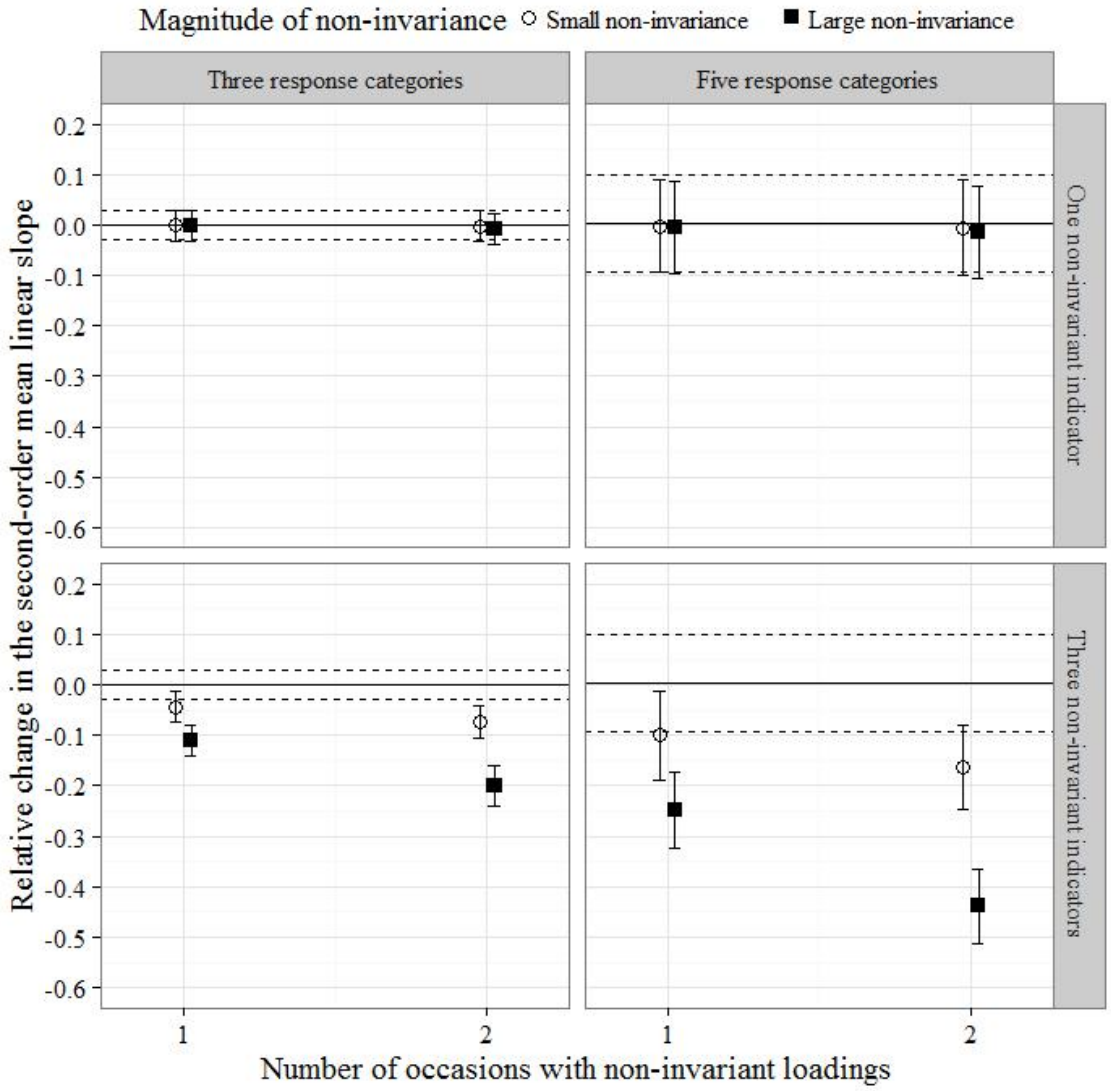

Figure 8. Mean relative change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal-theory 95% confidence interval of the relative change value in the corresponding baseline condition.

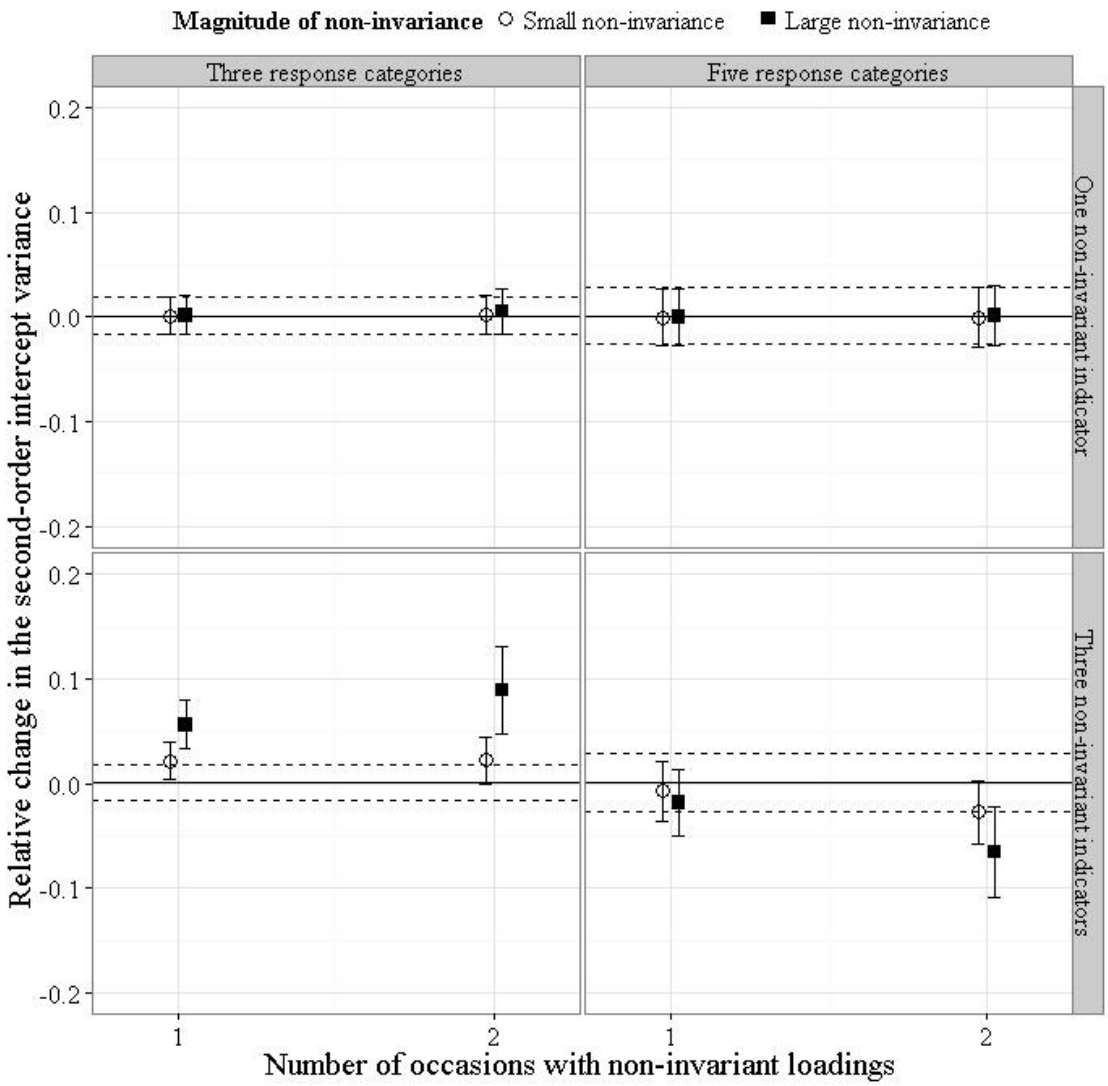

Figure 9. Mean relative change in the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

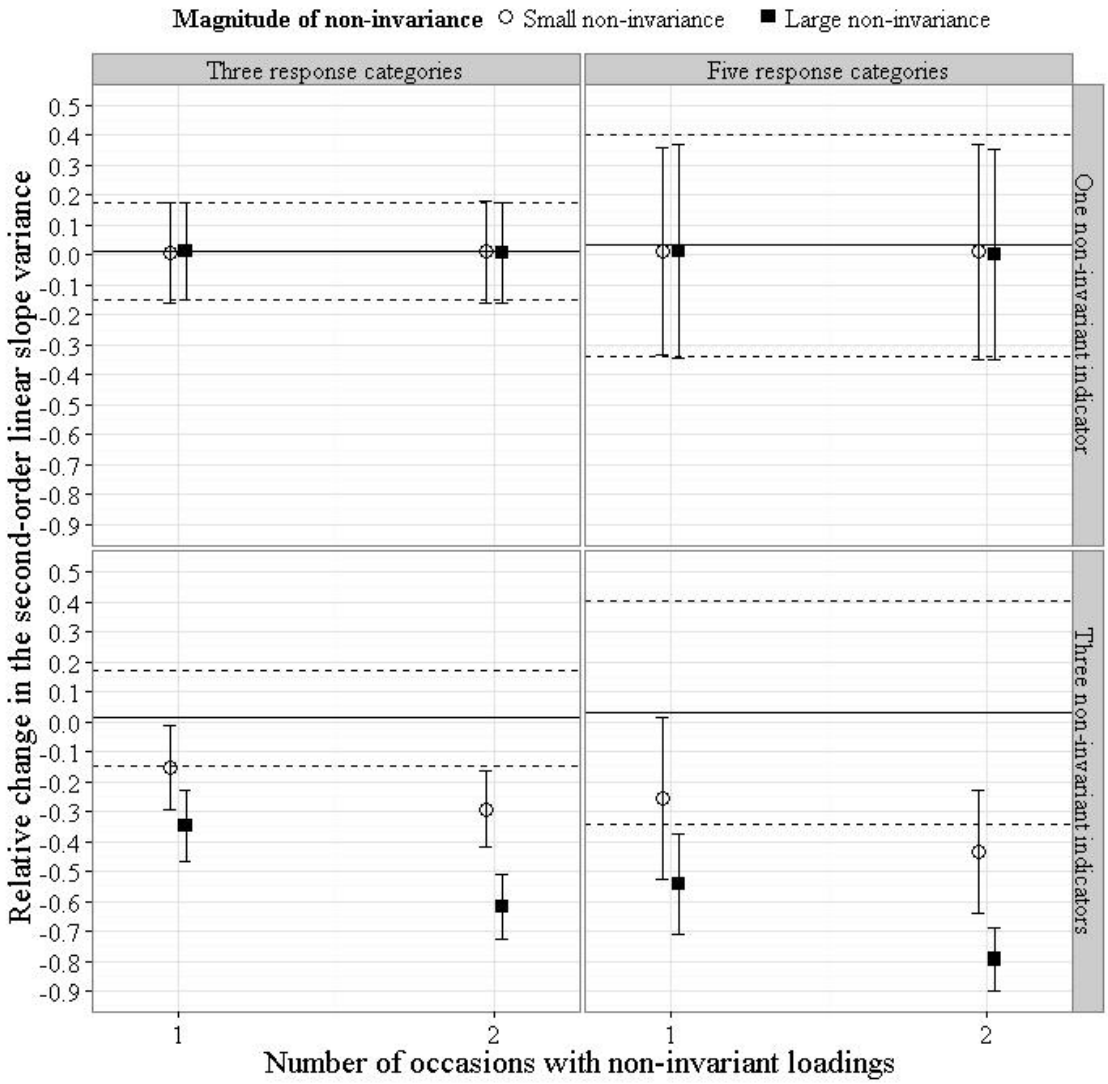

Figure 10. Mean relative change in the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

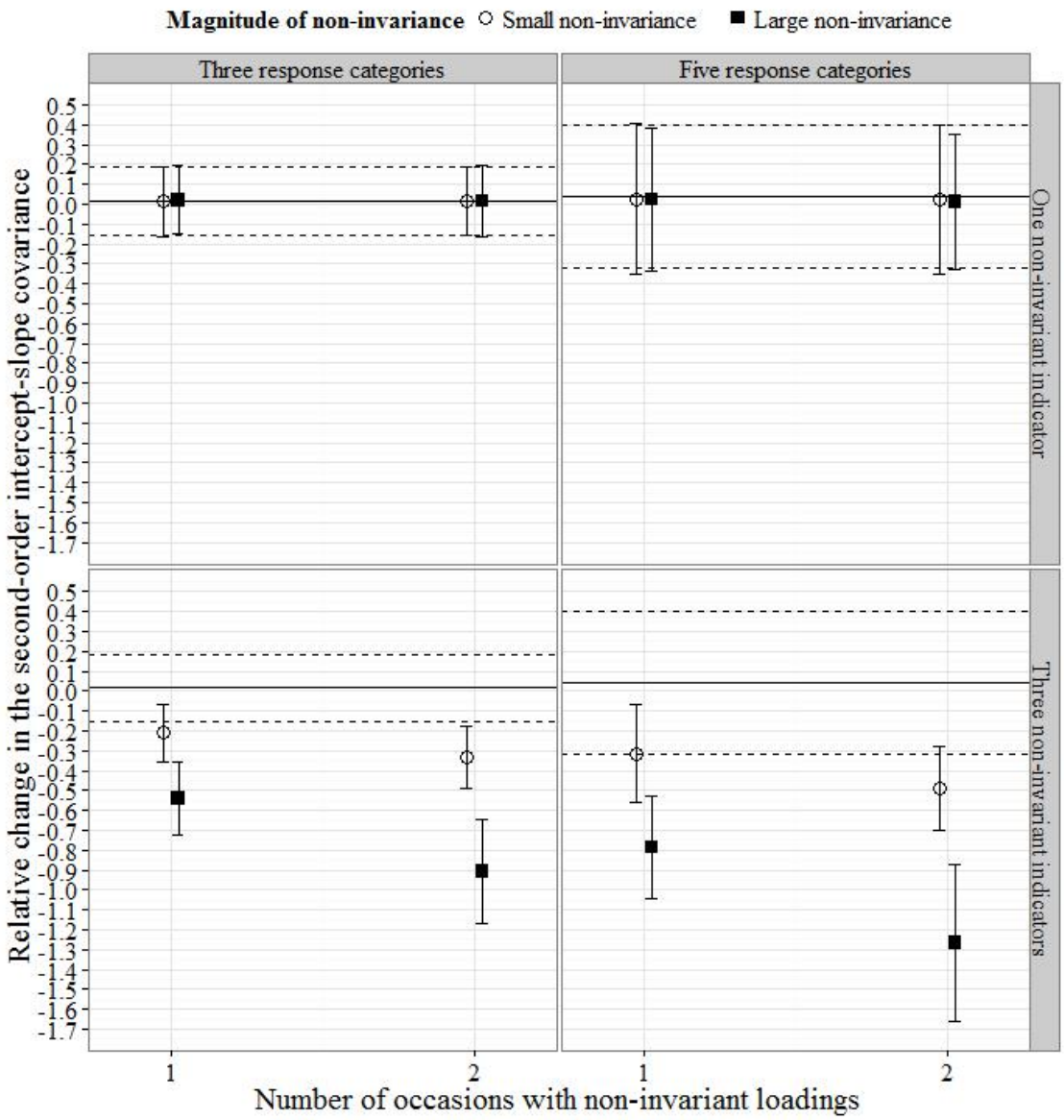

Figure 11. Mean relative change in the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

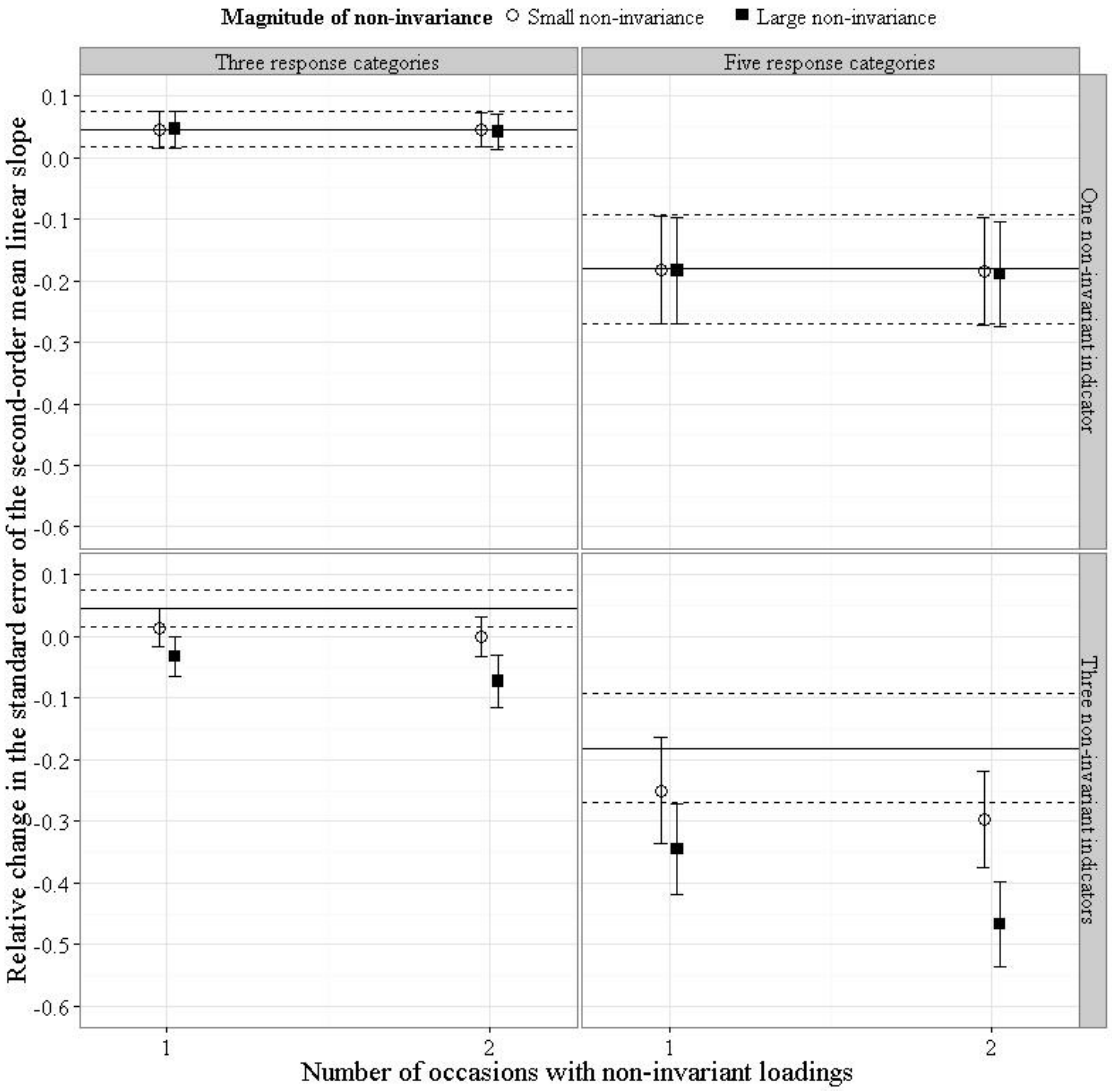

Figure 12. Mean relative change in the standard error of the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

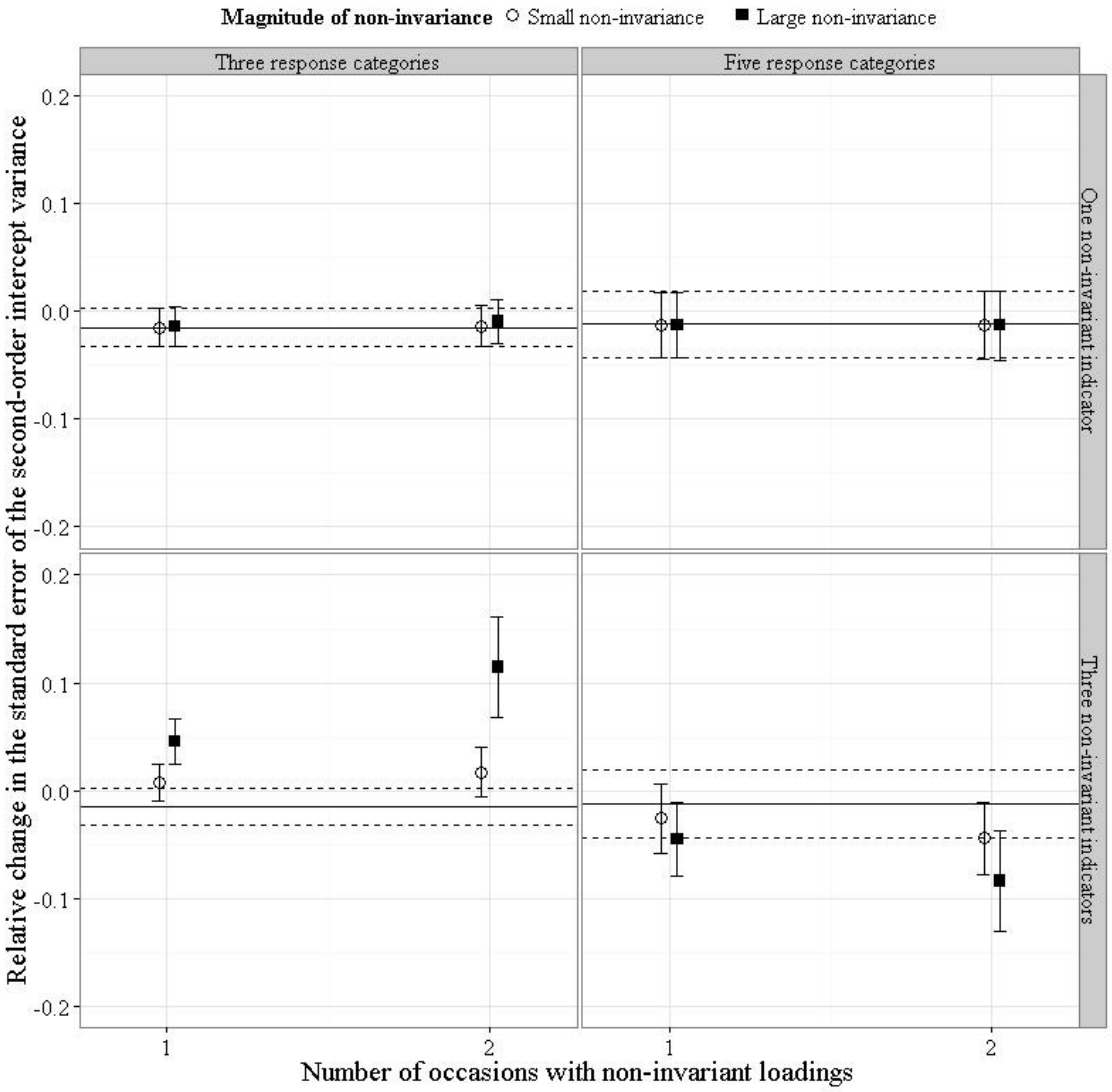

Figure 13. Mean relative change in the standard error of the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

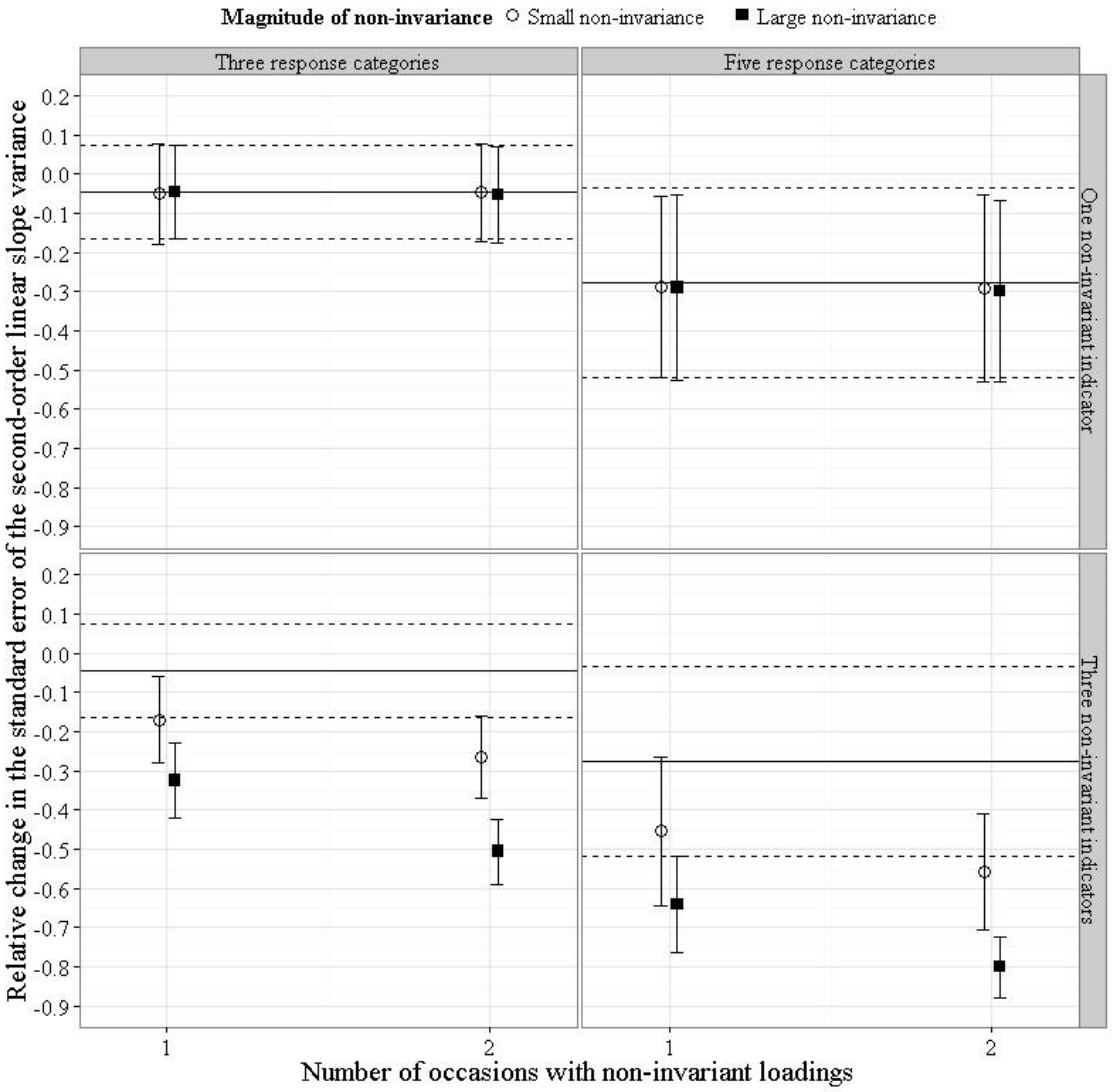

Figure 14. Mean relative change in the standard error of the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

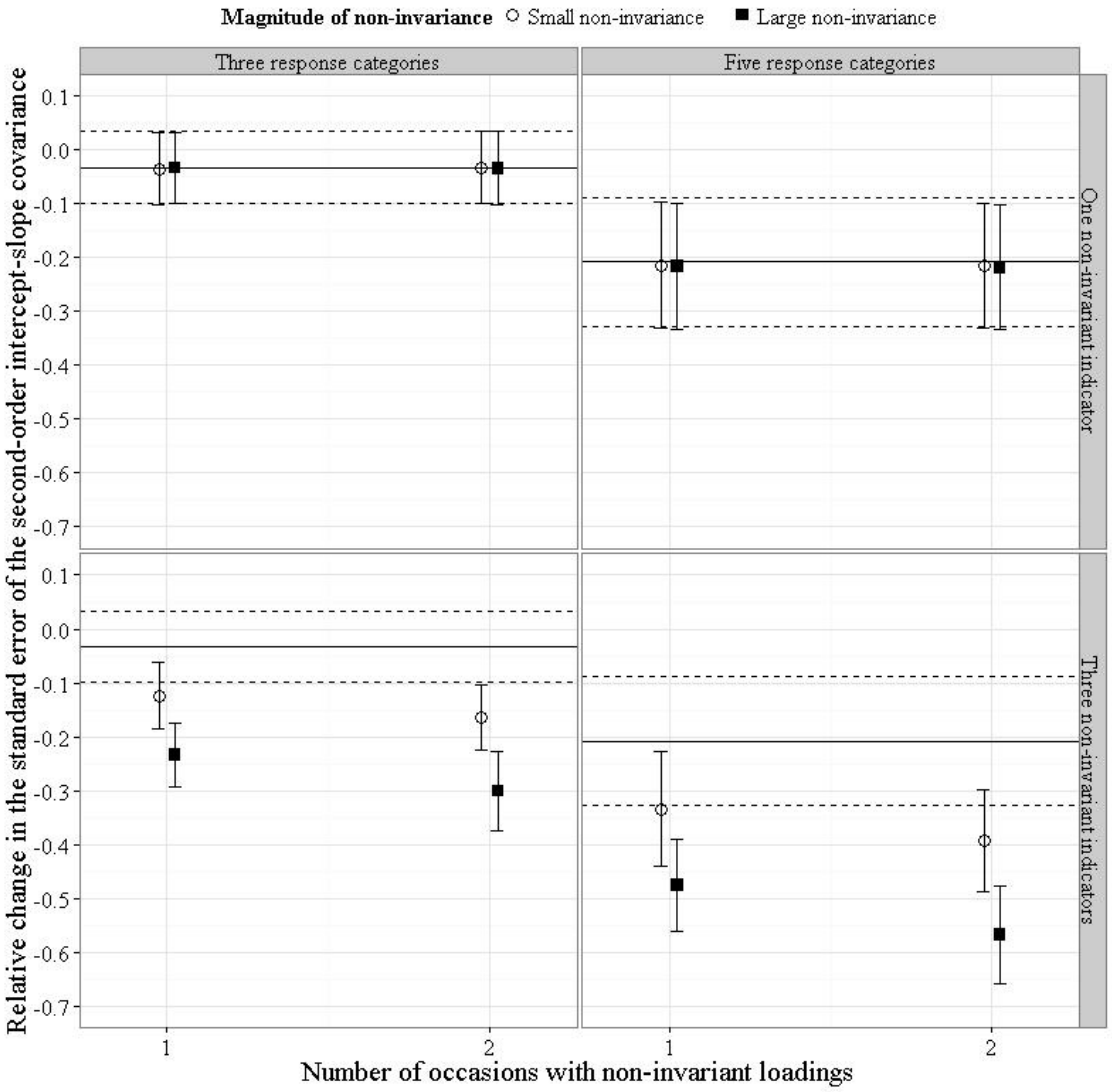

Figure 15. Mean relative change in the standard error of the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

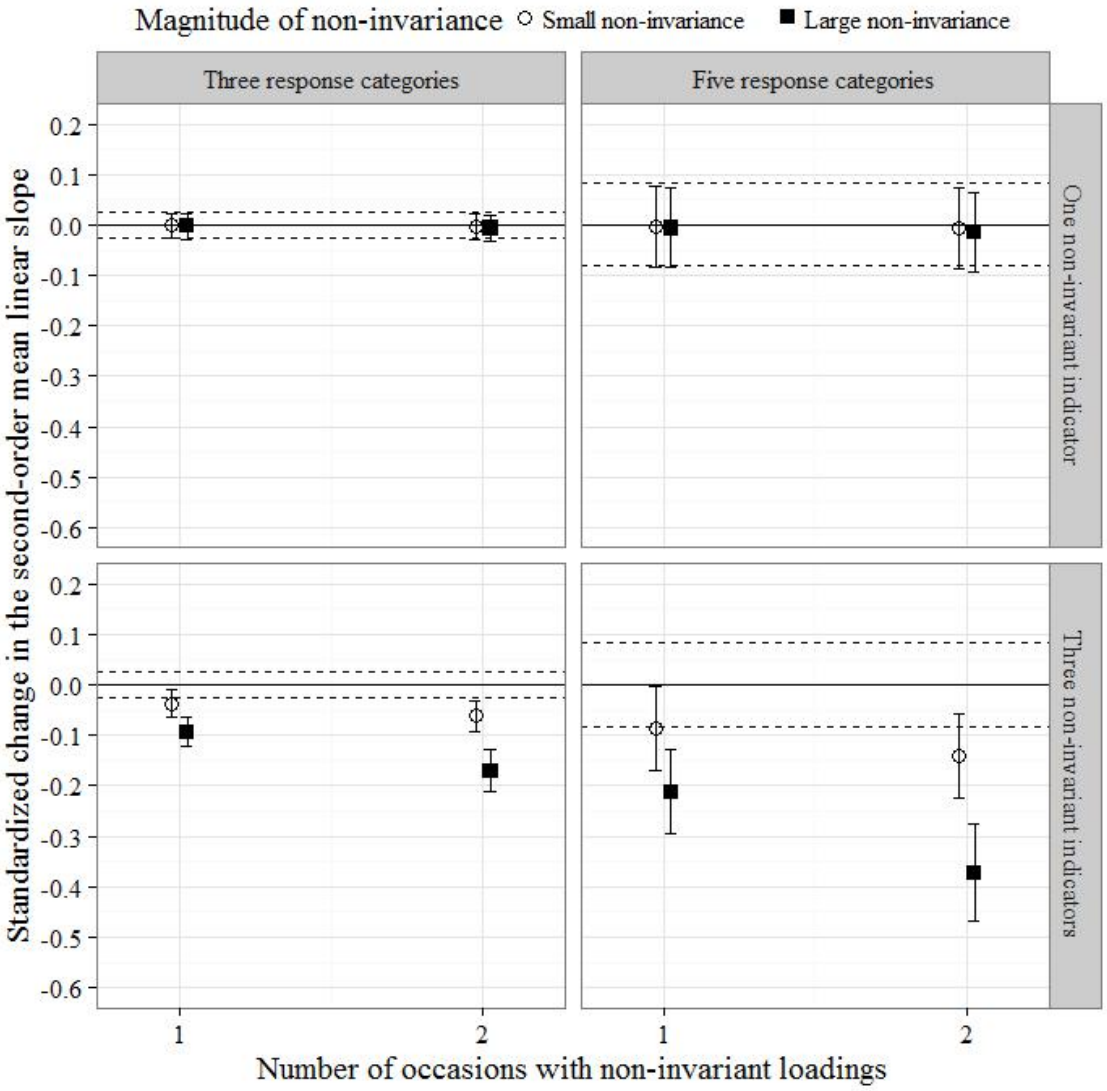

Figure 16. Mean standardized change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming configural invariance to the model incorrectly assuming loading invariance. *Note*: The solid horizontal line represents the mean standardized change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean standardized change value in the corresponding baseline condition.

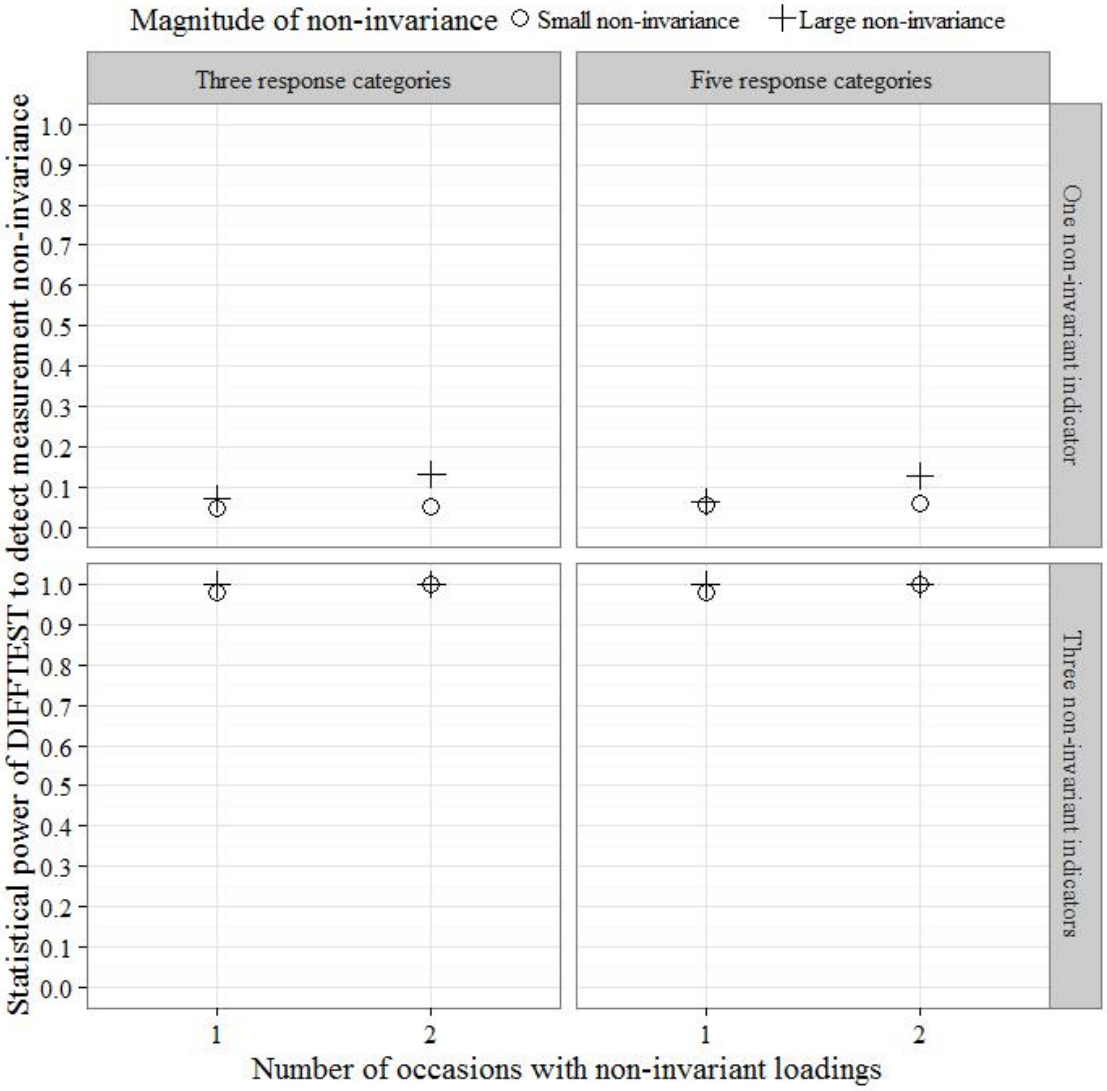

Figure 17. Statistical power of DIFFTEST to detect loading non-invariance, between the model correctly assuming configural invariance and the model incorrectly assuming loading invariance.

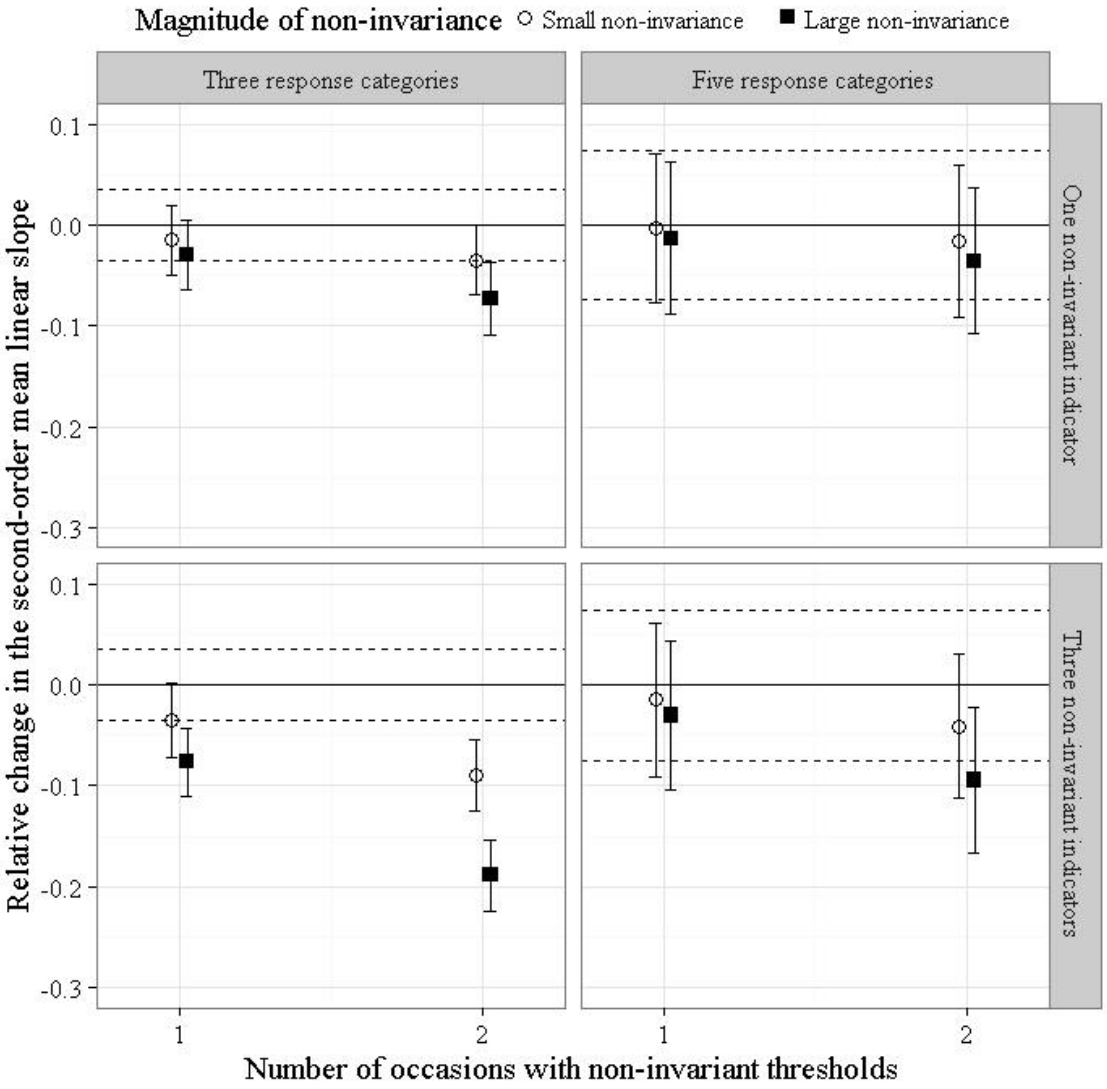

Figure 18. Mean relative change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal-theory 95% confidence interval of the relative change value in the corresponding baseline condition.

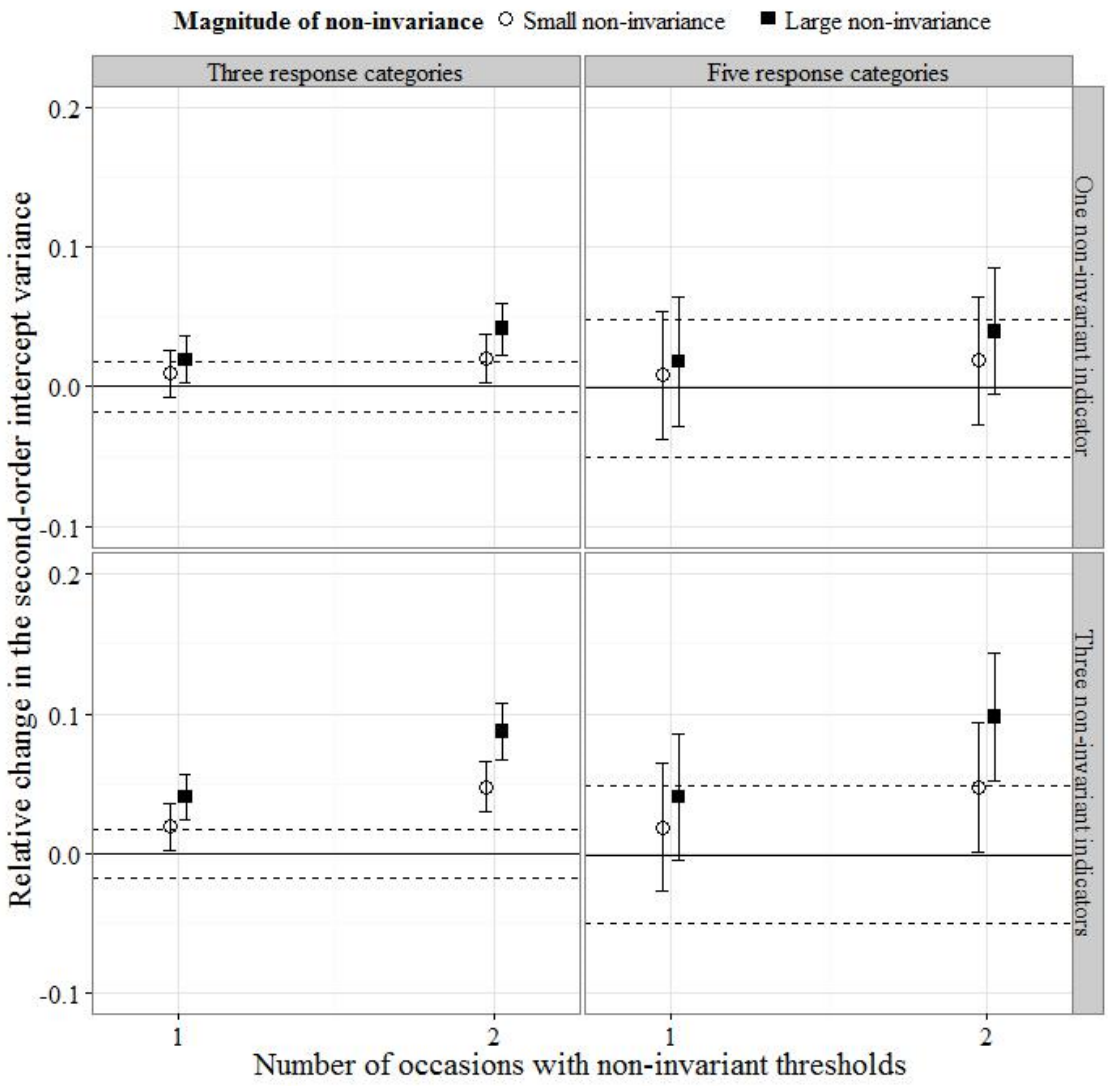

Figure 19. Mean relative change in the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

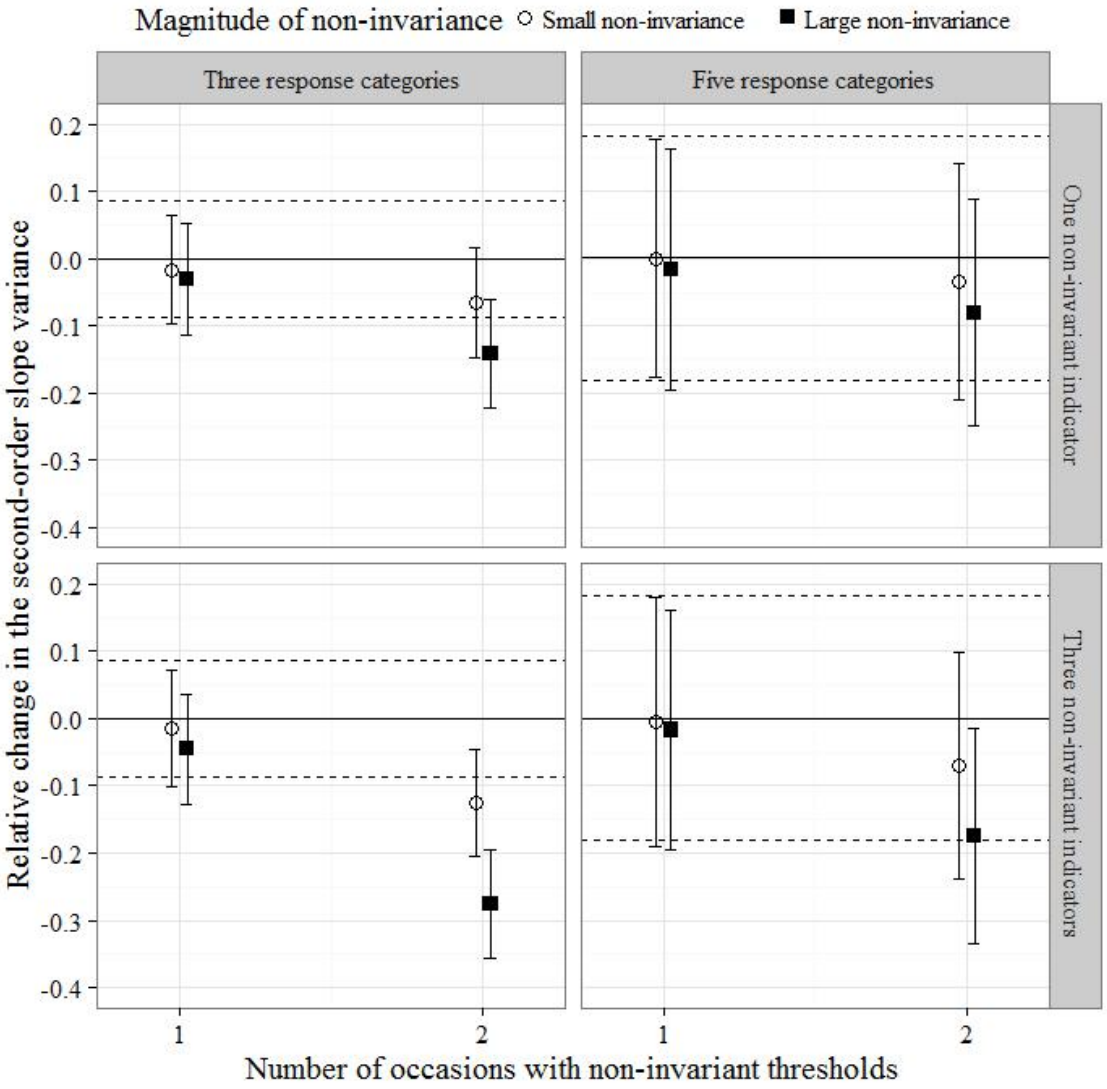

Figure 20. Mean relative change in the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

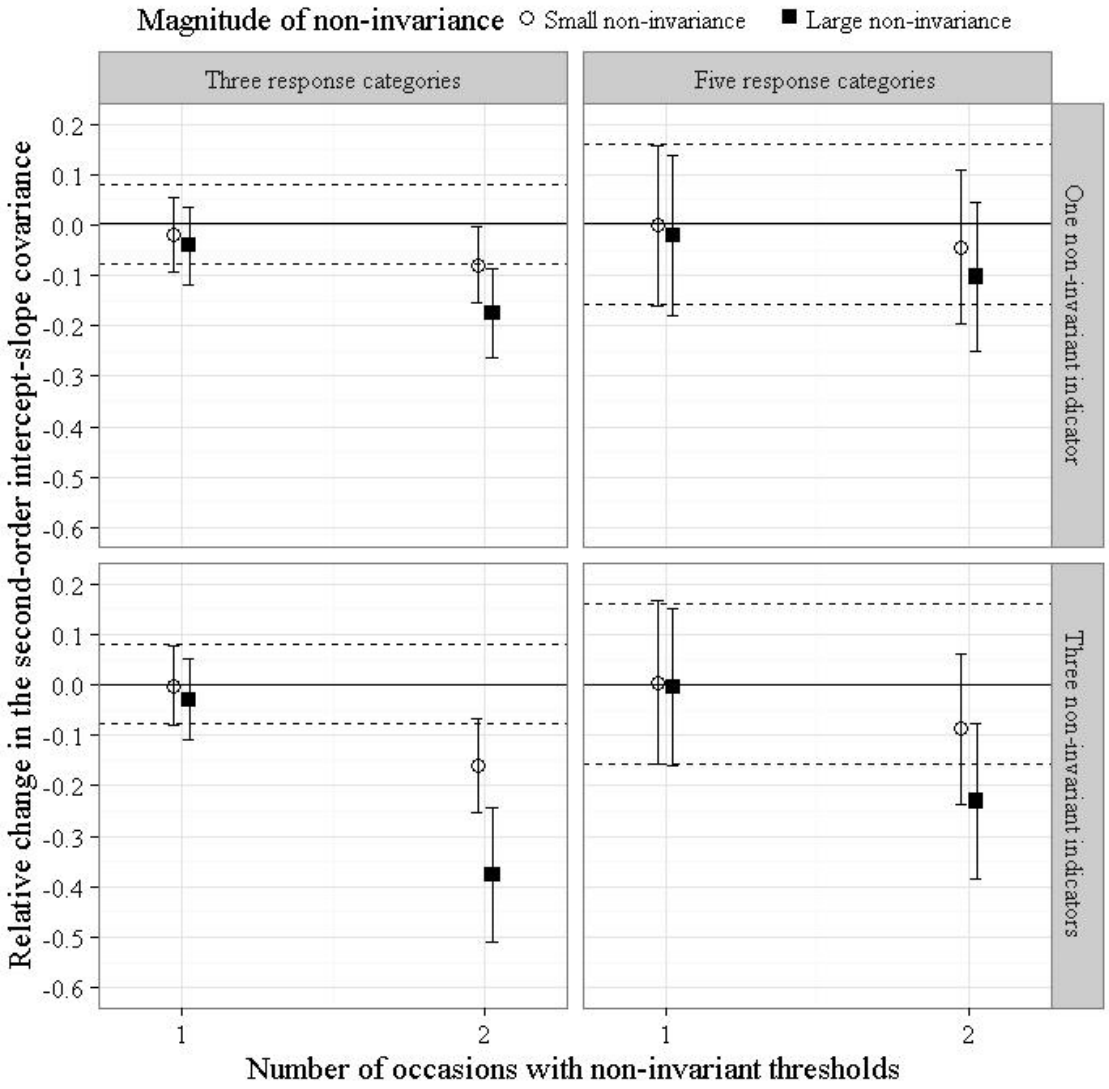

Figure 21. Mean relative change in the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

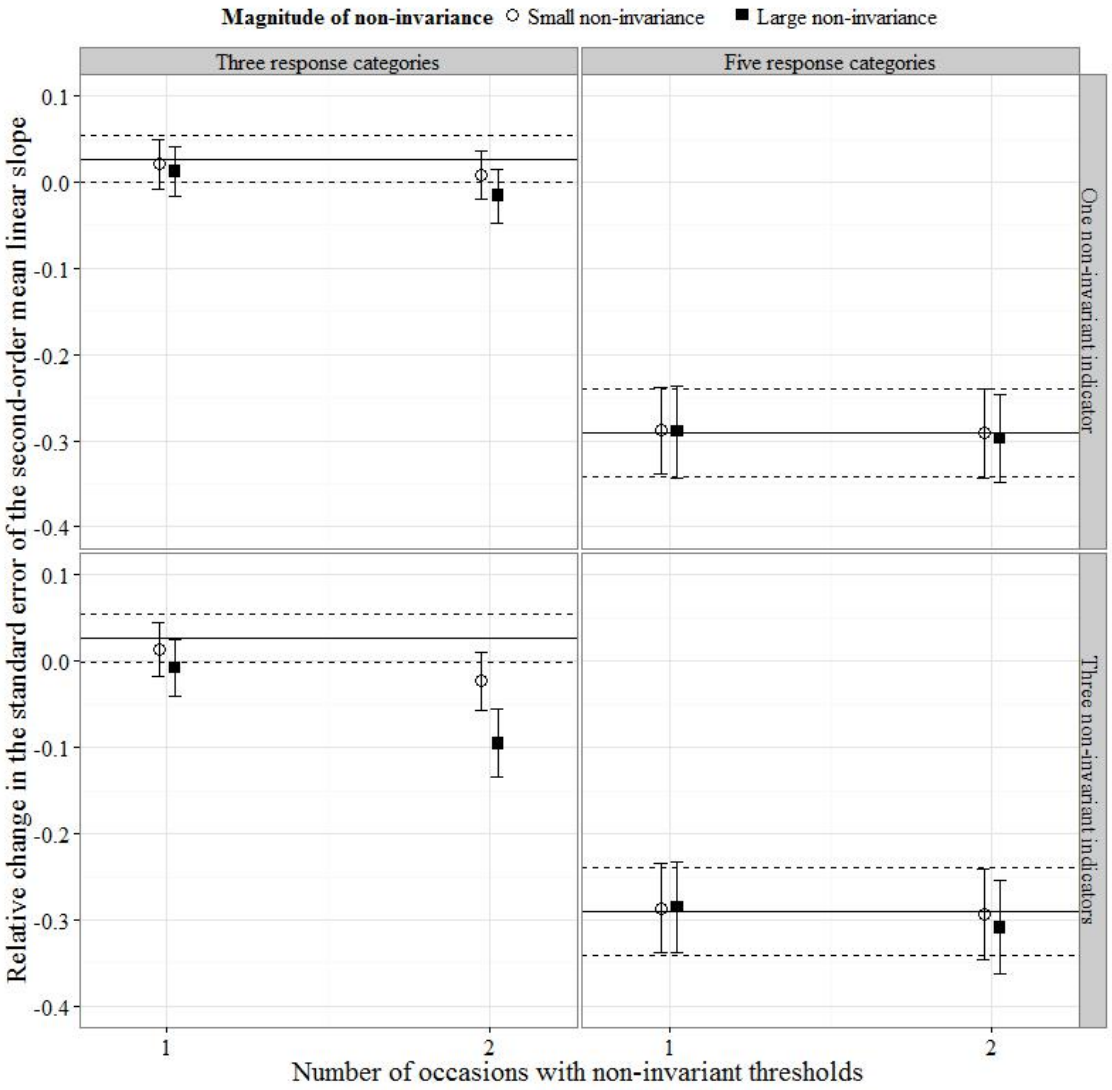

Figure 22. Mean relative change in the standard error of the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

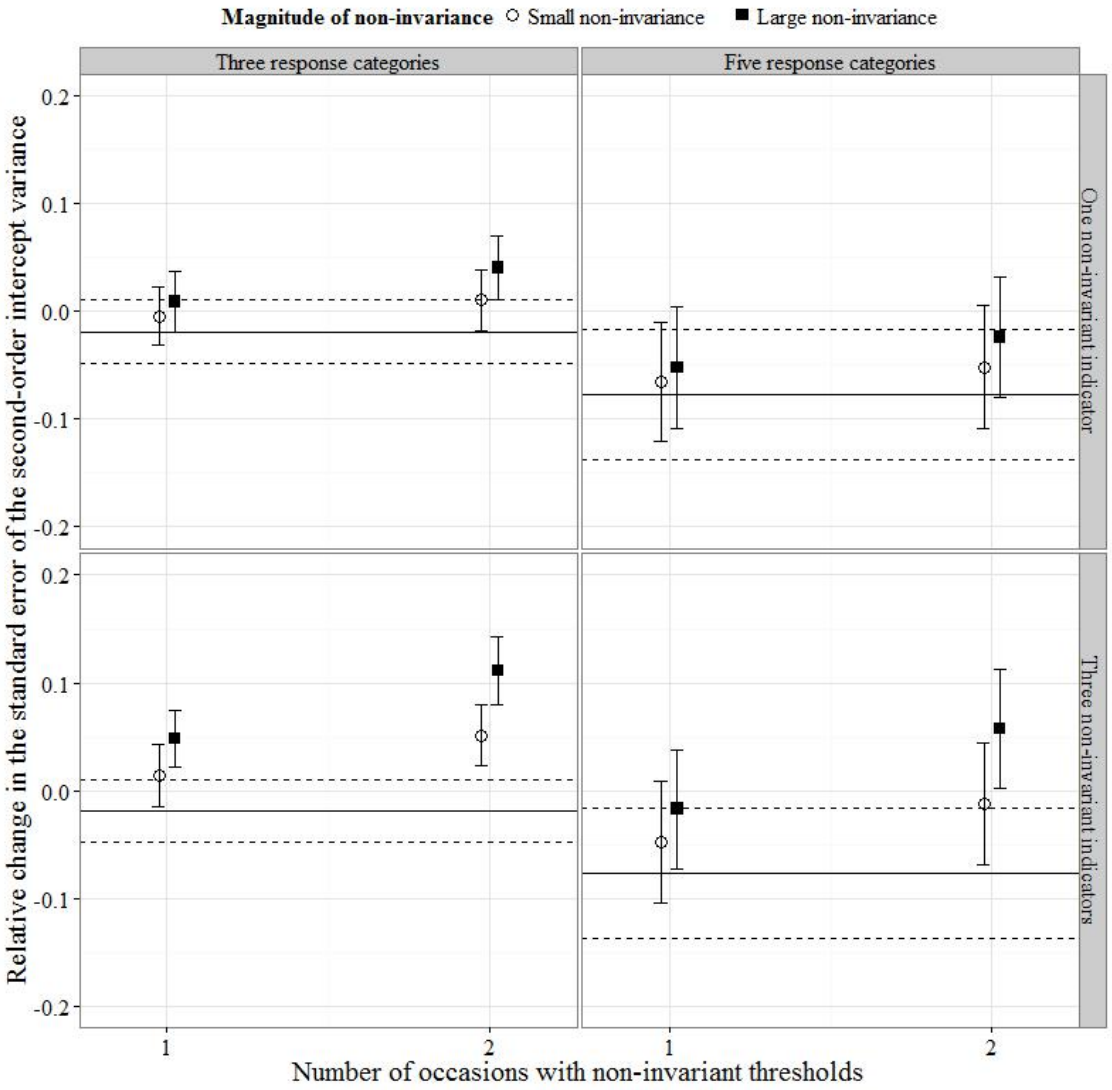

Figure 23. Mean relative change in the standard error of the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

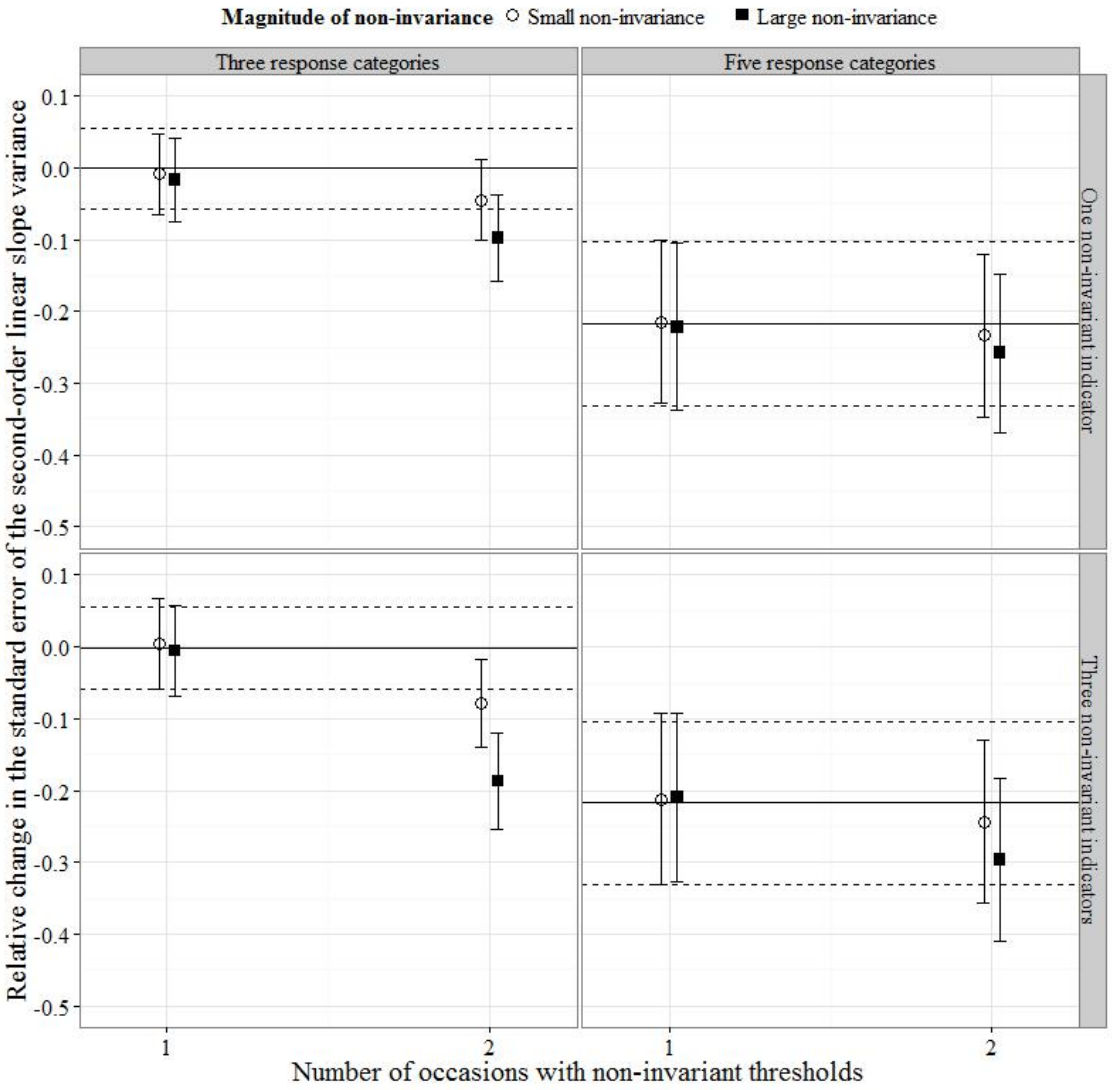

Figure 24. Mean relative change in the standard error of the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

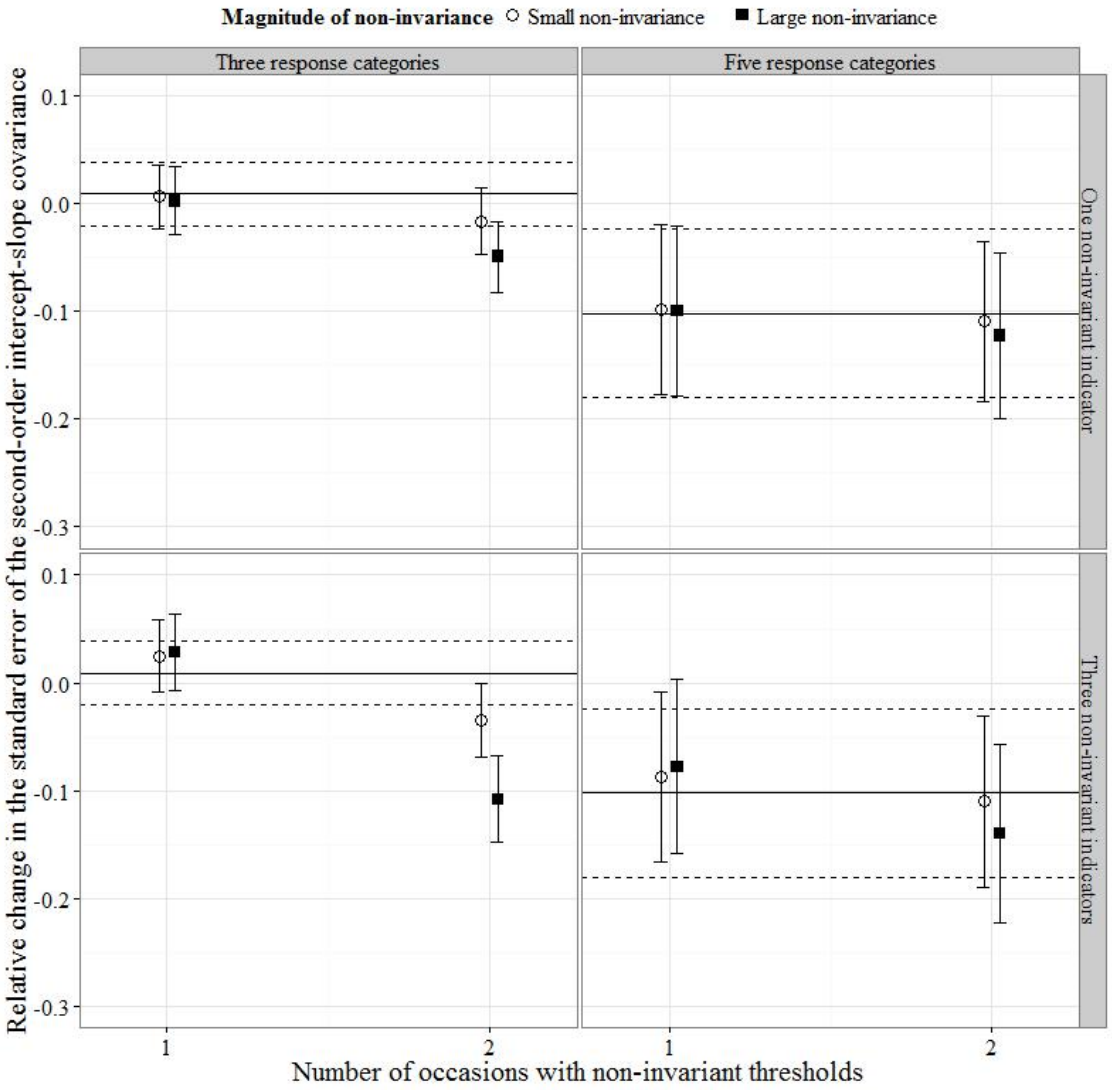

Figure 25. Mean relative change in the standard error of the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

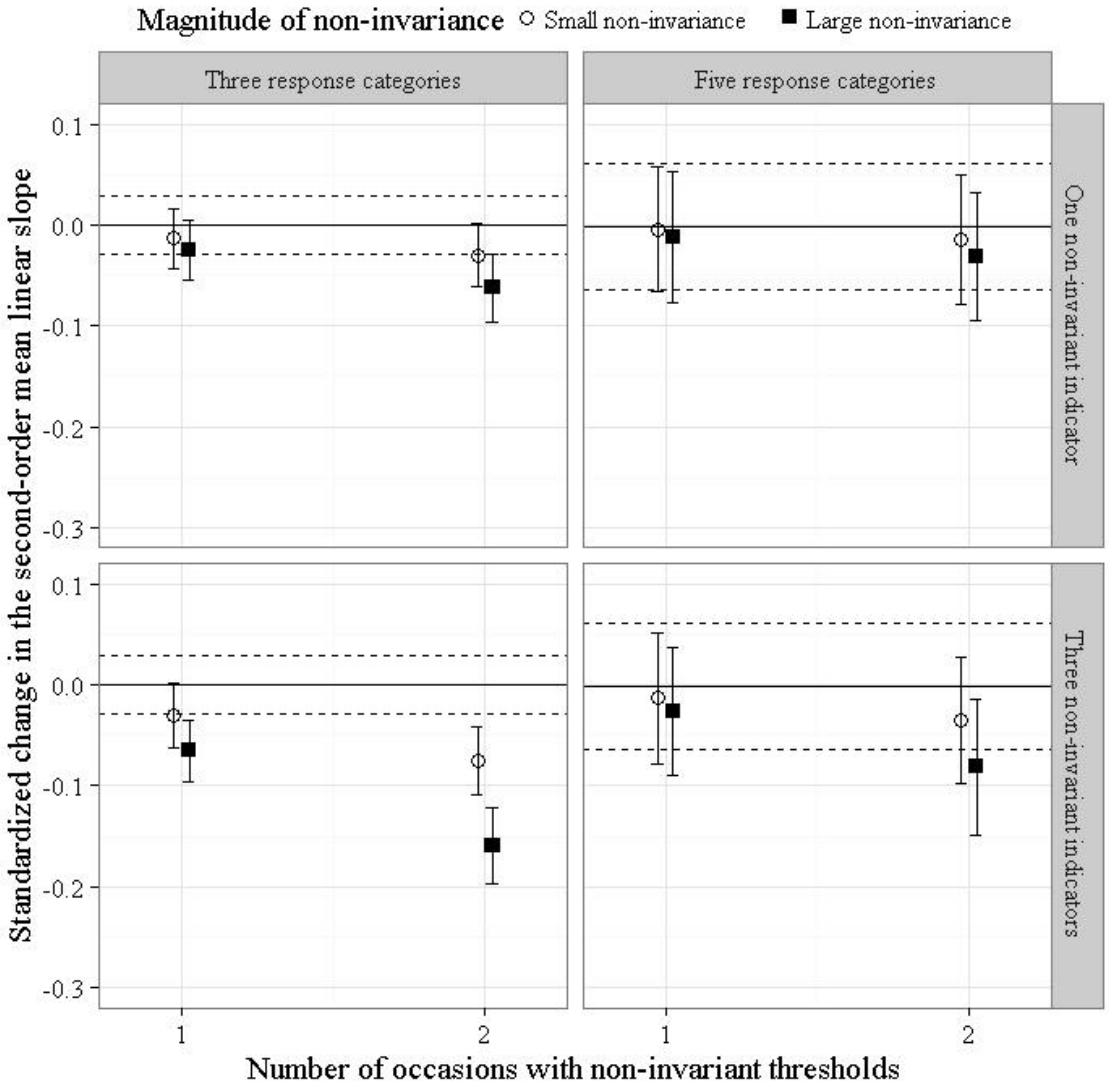

Figure 26. Mean standardized change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming loading invariance to the model incorrectly assuming threshold invariance. *Note*: The solid horizontal line represents the mean standardized change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean standardized change value in the corresponding baseline condition.

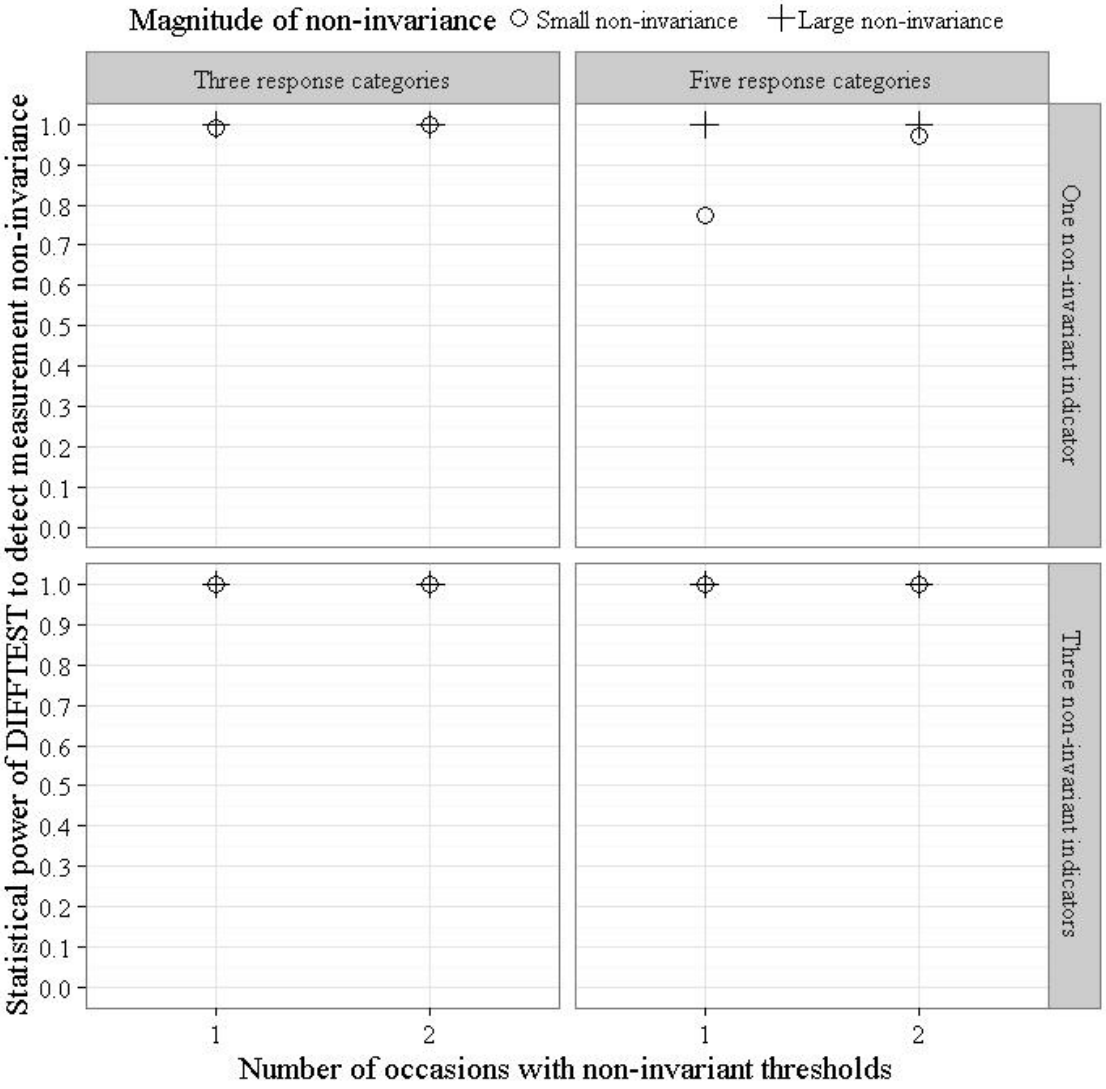

Figure 27. Statistical power of DIFFTEST to detect threshold non-invariance, between the model correctly assuming loading invariance and the model incorrectly assuming threshold invariance.

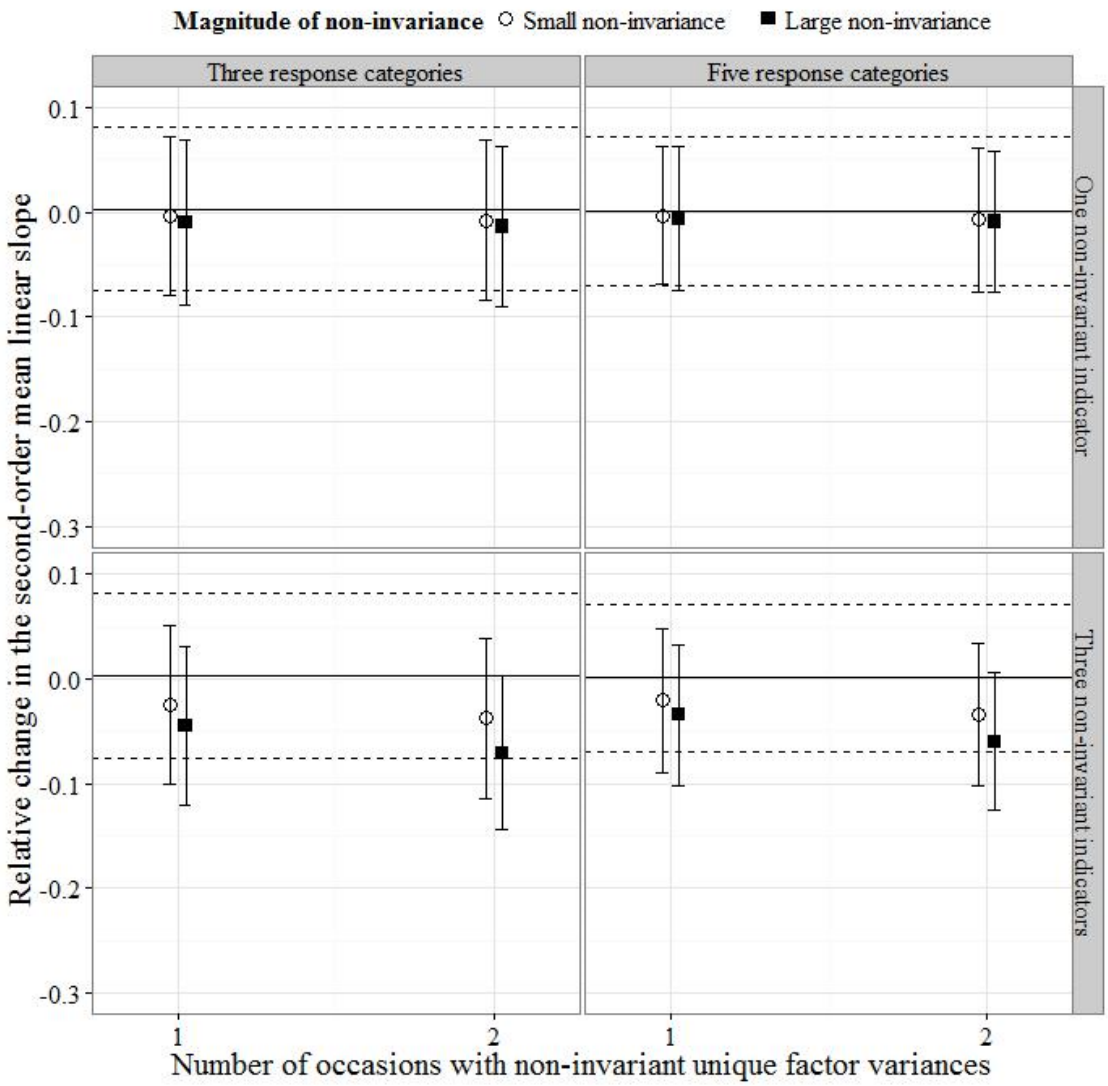

Figure 28. Mean relative change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal-theory 95% confidence interval of the relative change value in the corresponding baseline condition.

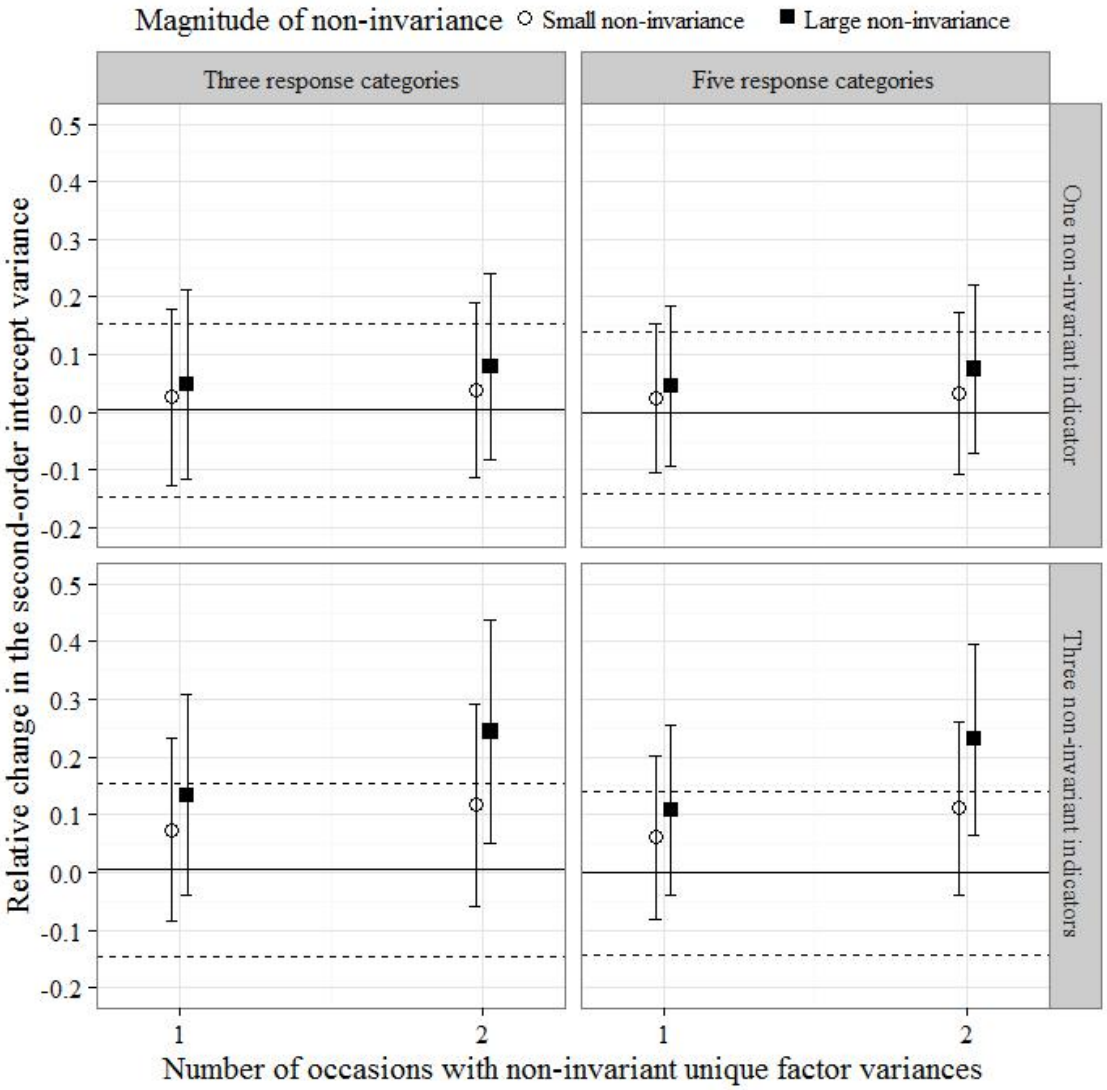

Figure 29. Mean relative change in the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note:* The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

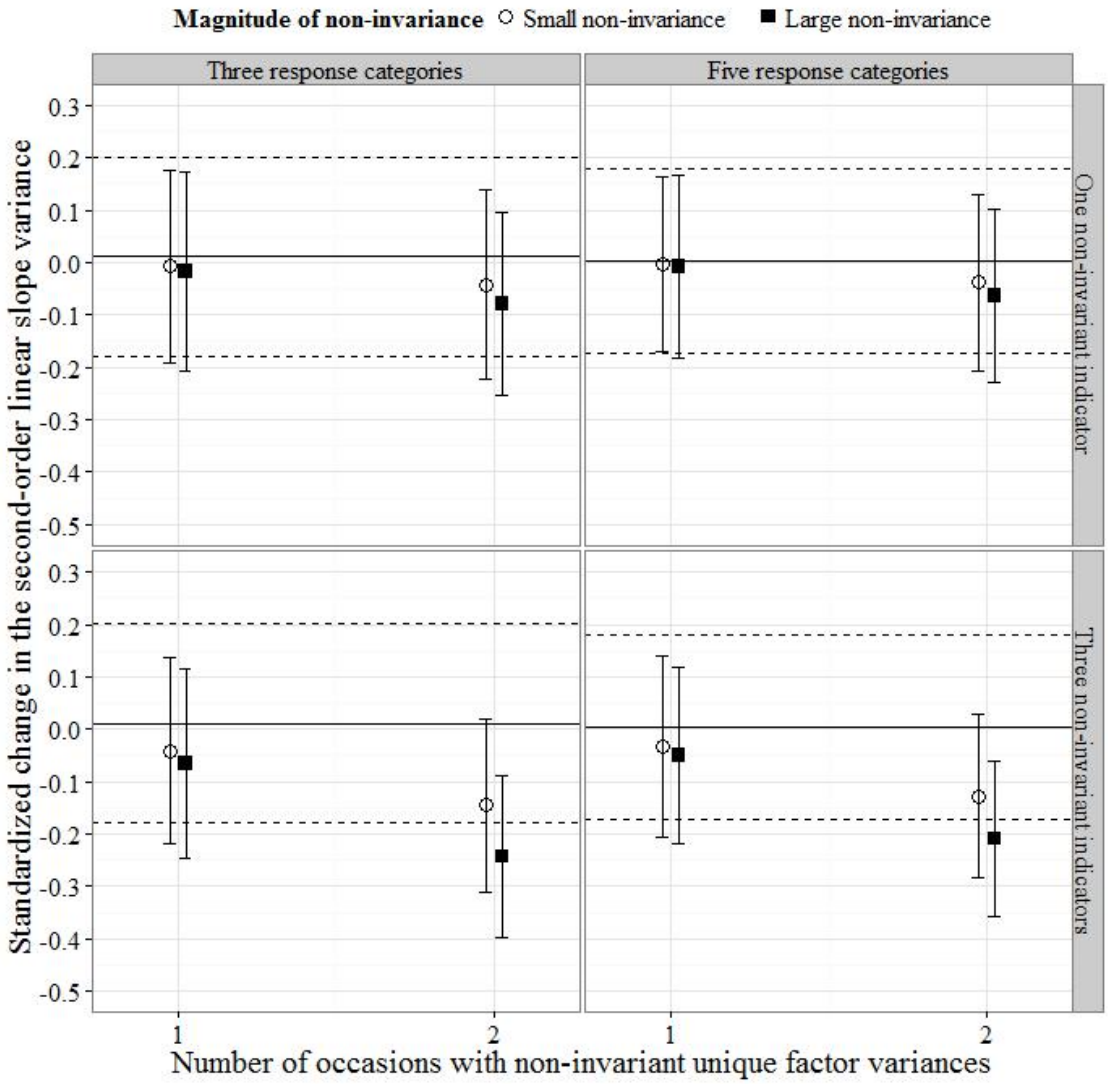

Figure 30. Mean relative change in the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

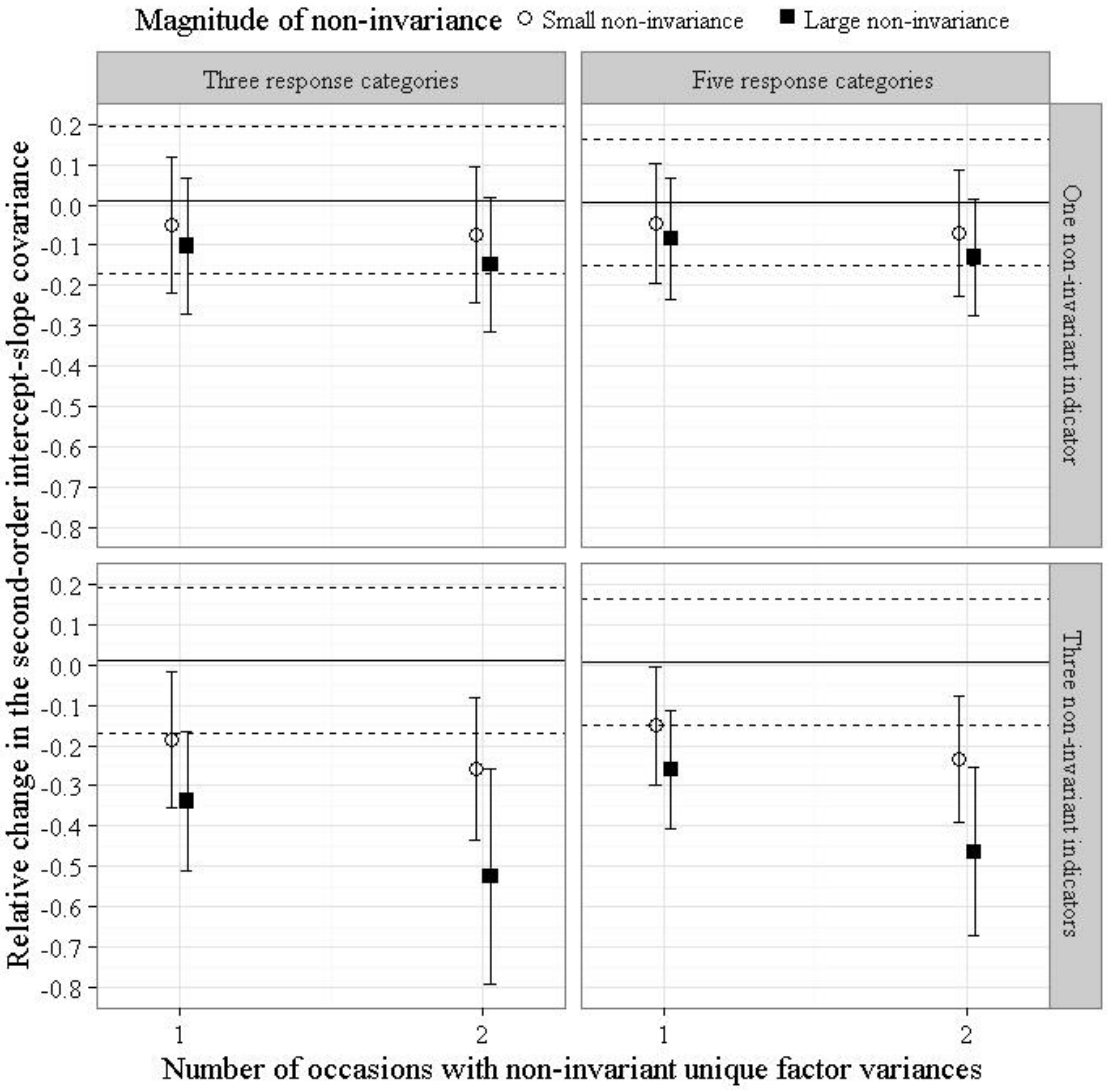

Figure 31. Mean relative change in the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

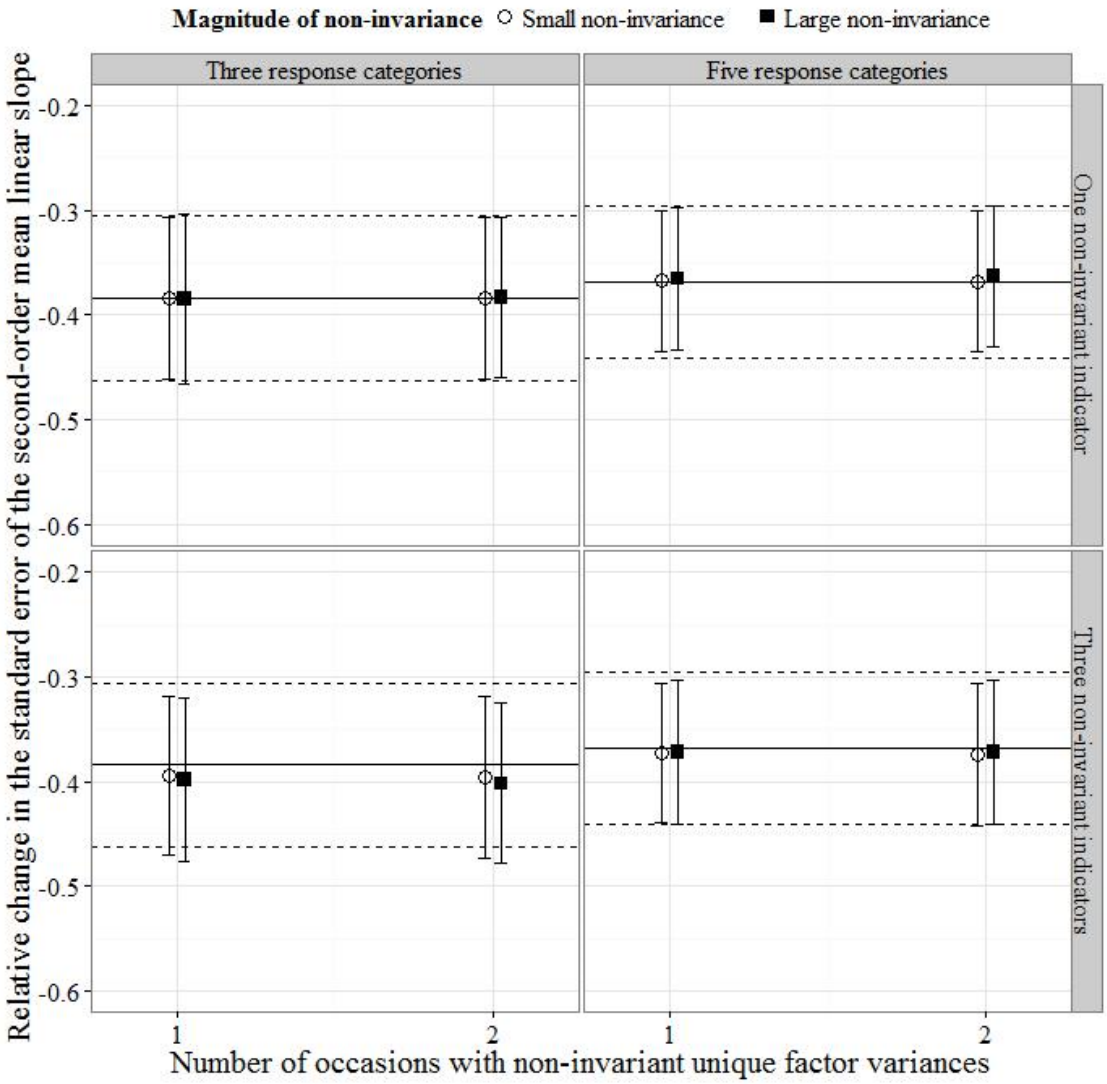

Figure 32. Mean relative change in the standard error of the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

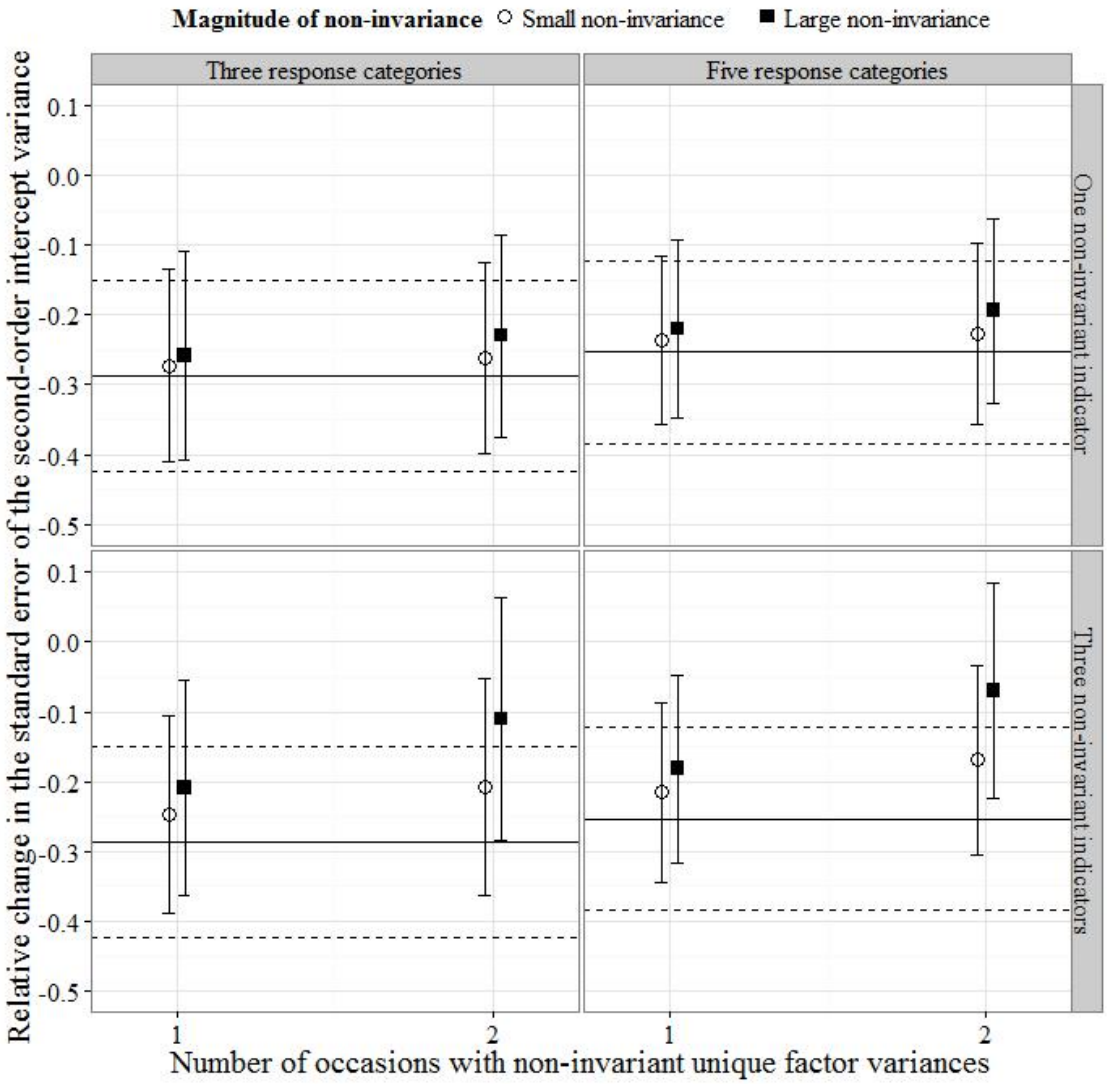

Figure 33. Mean relative change in the standard error of the second-order intercept variance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

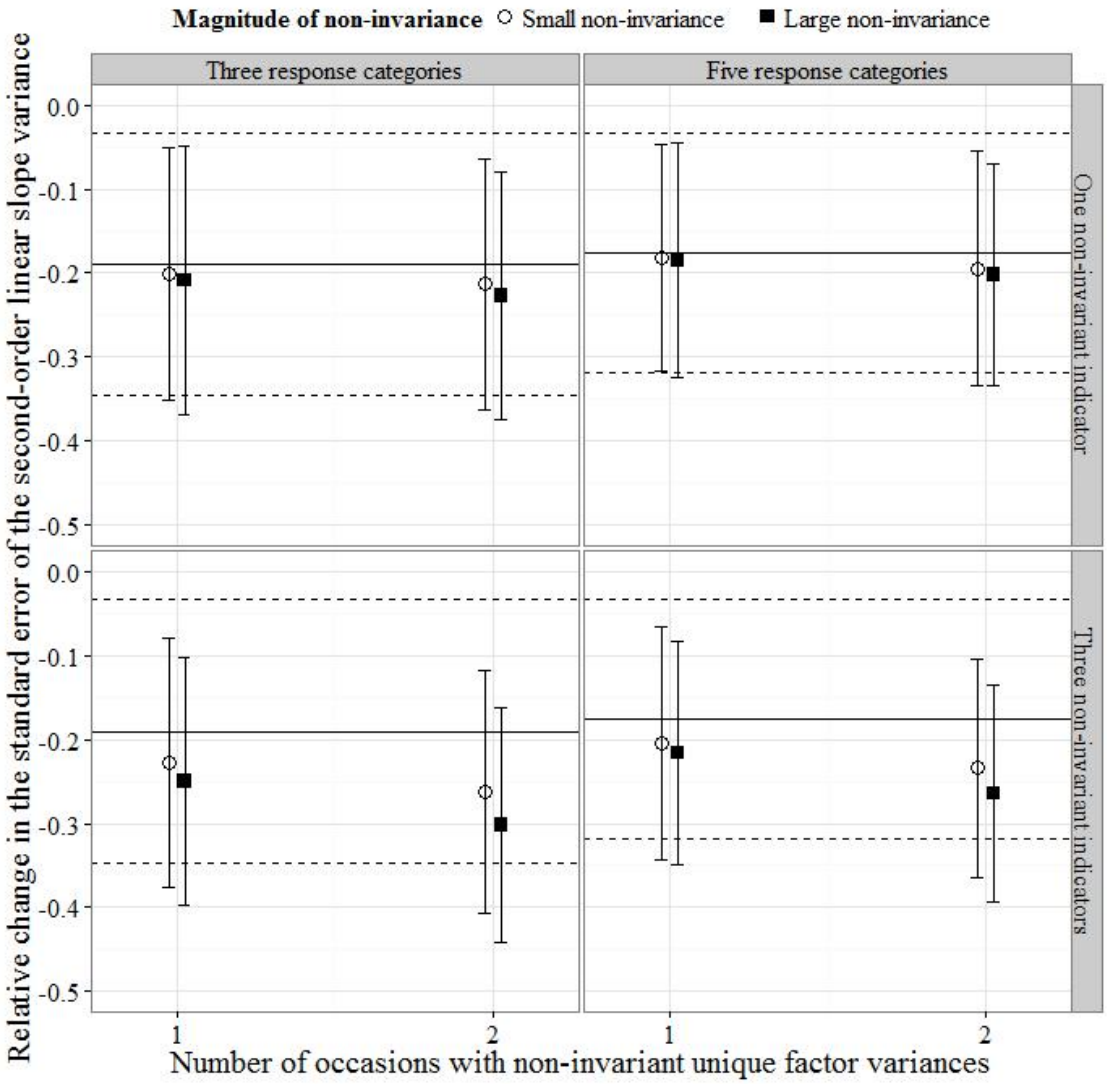

Figure 34. Mean relative change in the standard error of the second-order linear slope variance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

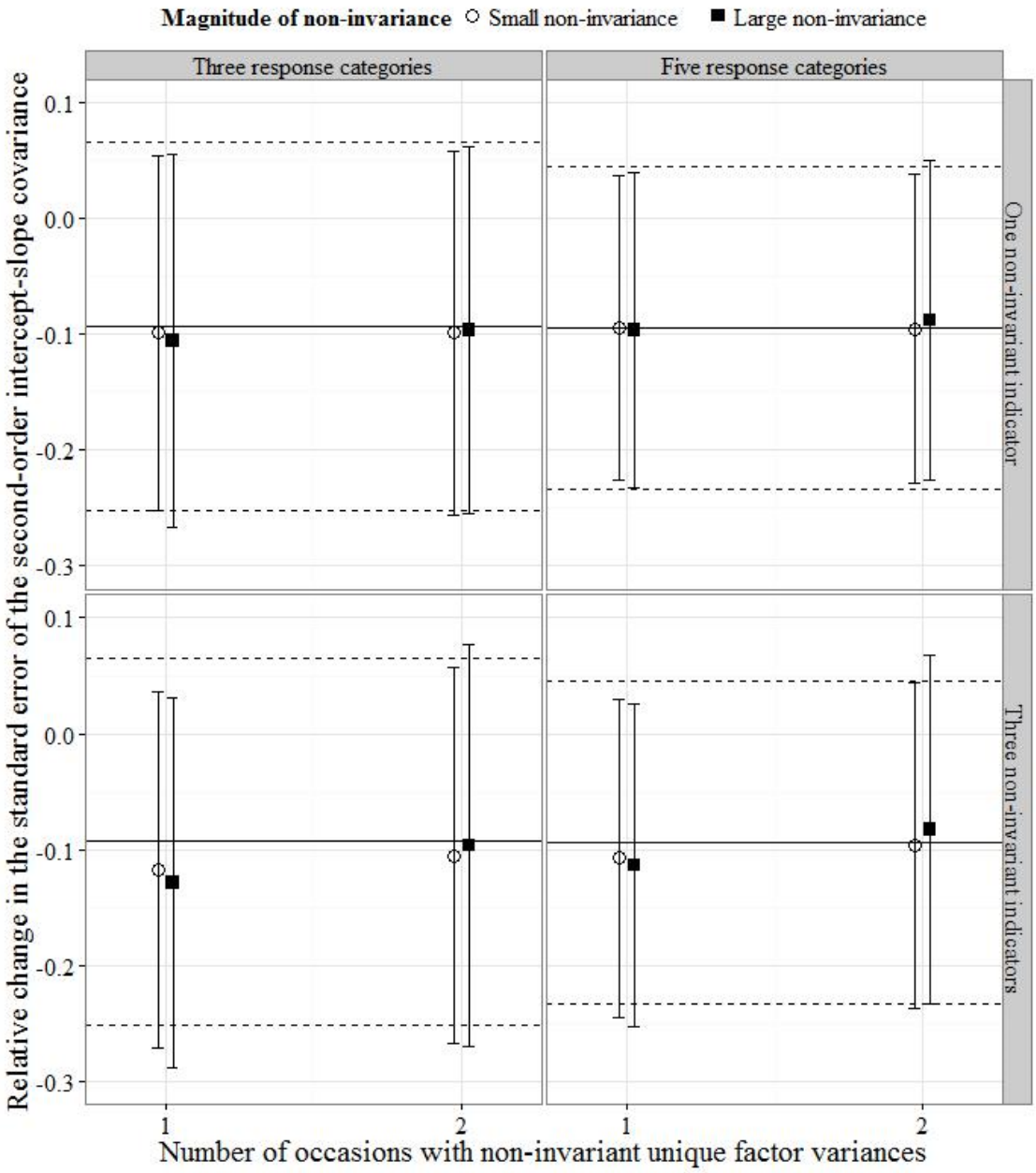

Figure 35. Mean relative change in the standard error of the second-order intercept-slope covariance with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean relative change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean relative change value in the corresponding baseline condition.

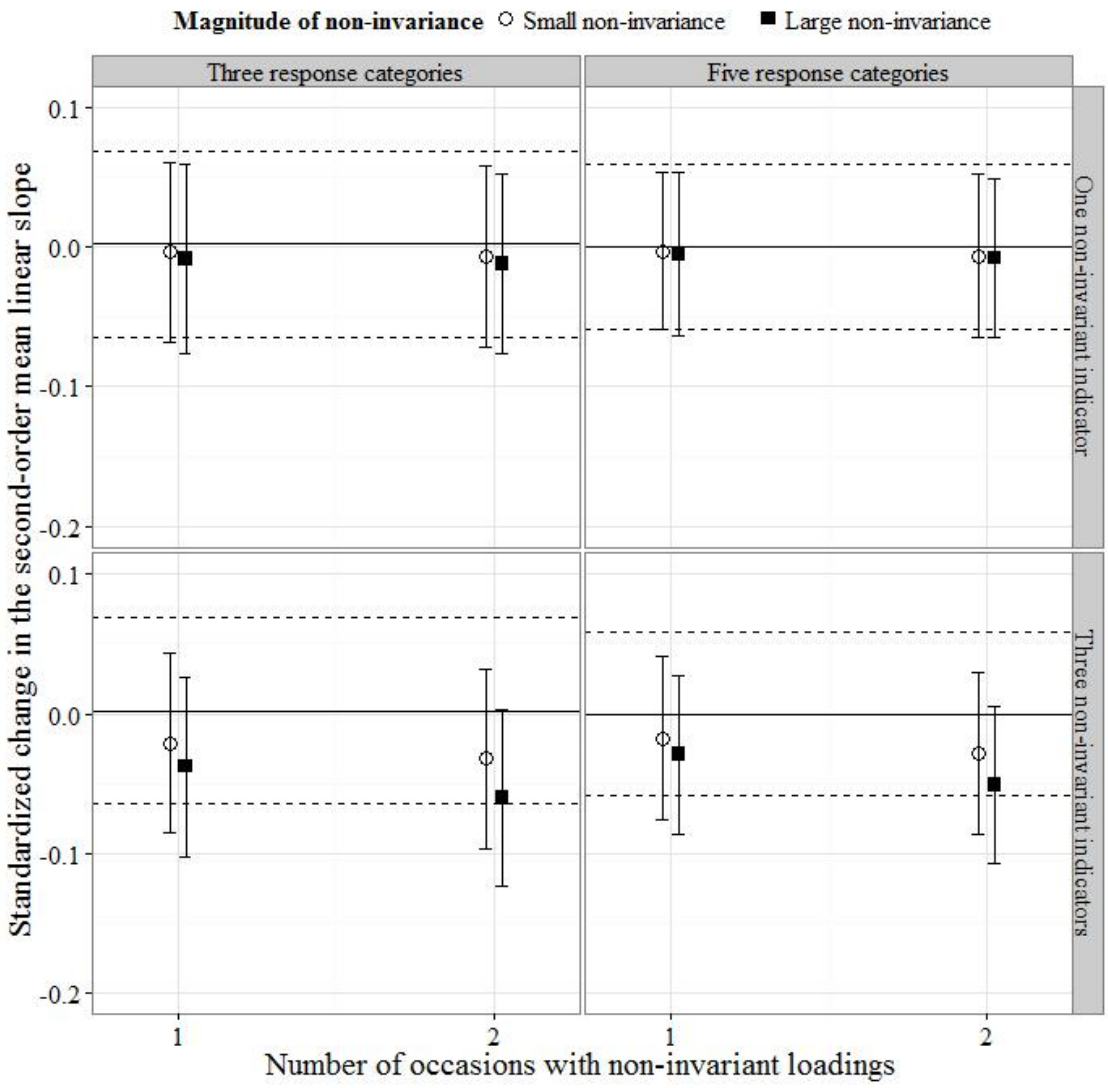

Figure 36. Mean standardized change in the second-order mean linear slope with normal-theory 95% confidence limits, from the model correctly assuming threshold invariance to the model incorrectly assuming unique factor invariance. *Note*: The solid horizontal line represents the mean standardized change value in the corresponding baseline condition. The dashed horizontal lines represent the upper and lower limits of the normal theory 95% confidence interval of the mean standardized change value in the corresponding baseline condition.

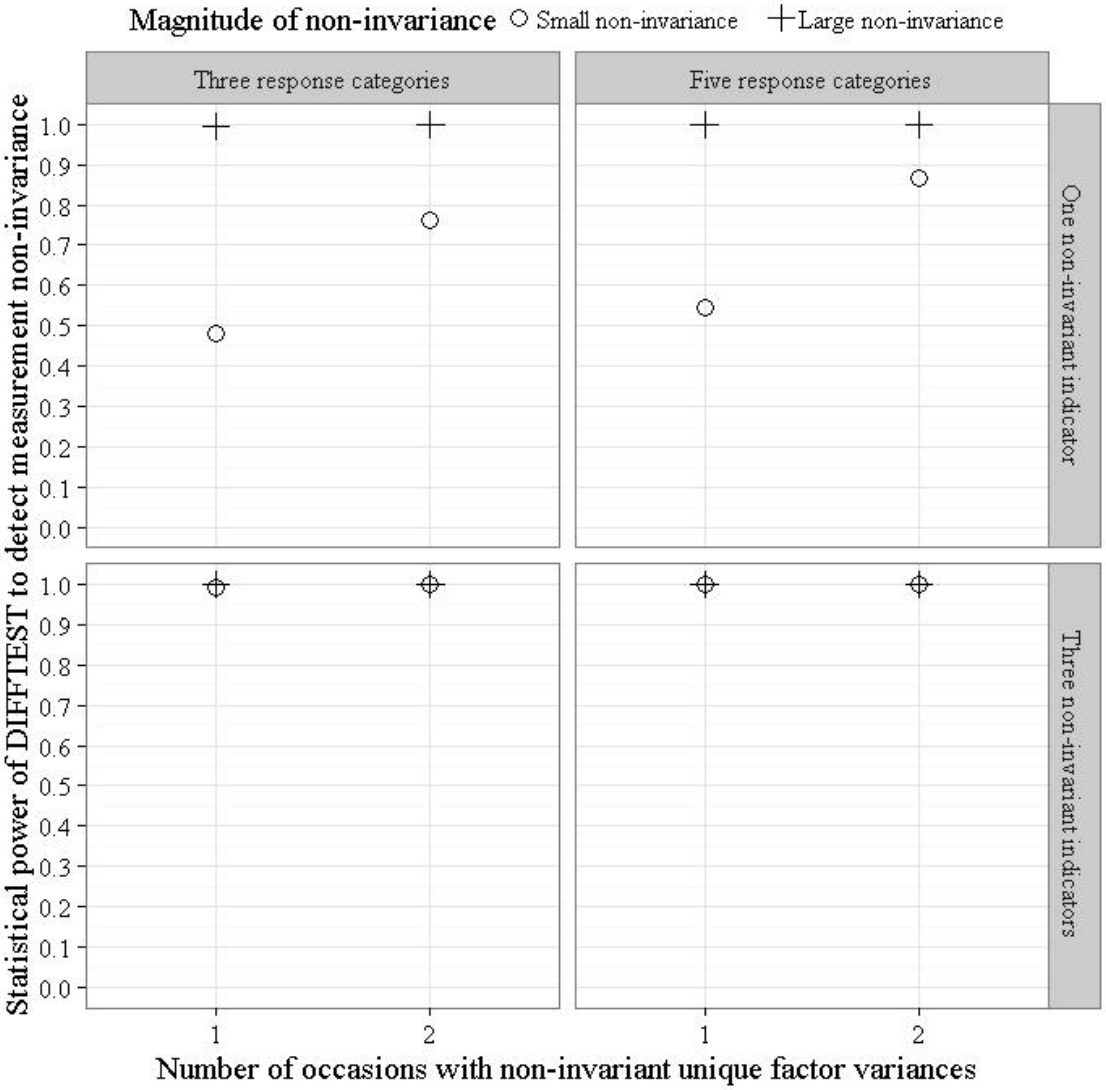

Figure 37. Statistical power of DIFFTEST to detect unique factor non-invariance, between the model correctly assuming threshold invariance and the model incorrectly assuming unique factor invariance.

REFERENCES

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186-203. DOI: 10.1207/s15328007sem1302_2

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective* (Vol. 467). Hoboken, NJ: John Wiley & Sons.

Brown, M. B., & Bendetti, J. K. (1977). On the mean and variance of the tetrachoric correlation coefficient. *Psychometrika*, *42*(3), 347–355. DOI: 10.1007/BF02293655

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). Mahwah, NJ: Erlbaum.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling*, *9*(3), 327-346. DOI: 10.1207/S15328007SEM0903_2

DiStefano, C., & Hess, B. (2005). Using confirmatory factor analysis for construct validation: An empirical review. *Journal of Psychoeducational Assessment*, *23*(3), 225-241. DOI: 10.1177/073428290502300303

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling*, *21*(3), 425-438. DOI:10.1080/10705511.2014.915373

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

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466-491. DOI: 10.1037/1082-989X.9.4.466

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275-299. DOI: 10.1037/a0015825

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling, 16*(4), 625-641. DOI: 10.1080/10705510903203573

Gonzalez-Roma, V., Hernandez, A., & Gomez-Benito, J. (2006). Power and Type I error of the mean and covariance structure analysis model for detecting differential item functioning in graded response items. *Multivariate Behavioral Research, 41*(1), 29-53. DOI: 10.1207/s15327906mbr4101_3

Grimm, K. J. & Liu, Y. (2016). Residual structures in growth models with ordinal outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*. DOI: 10.1080/10705511.2015.1103192

Grimm, K. J., Ram, N., & Estabrook, R. (in press). *Growth modeling: Structural equation and multilevel modeling approaches*. New York: Guilford Press.

Hoshino, T. & Bentler, P. M. (2011). Bias in factor score regression and a simple solution. In A.R. de Leon & K.C. Chough (Eds.), *Analysis of mixed data: Methods & applications*. New York: Taylor & Francis.

Khoo, S. T., West, S. G., Wu, W., & Kwok, O. M. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 301-317). Washington, D.C.: American Psychological Association.

Kim, E. S., & Willson, V. L. (2014a). Measurement invariance across groups in latent growth Modeling. *Structural Equation Modeling, 21*(3), 408-424. DOI: 10.1080/10705511.2014.915374

Kim, E. S., & Willson, V. L. (2014b). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model. *Structural Equation Modeling, 21*(4), 566-576. DOI: 10.1080/10705511.2014.919821

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling, 18*(2), 212-228. DOI: 10.1080/10705511.2011.557337

Koran, J., & Hancock, G. R. (2010). Using fixed thresholds with grouped data in structural equation modeling. *Structural Equation Modeling, 17*(4), 590-604. DOI: 10.1080/10705511.2010.510047

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review, 23*(4), 418-444. DOI: 10.1177/0193841X9902300404

Kuha, J., & Moustaki, I. (2015). Nonequivalence of measurement in latent variable modeling of multigroup data: A sensitivity analysis. *Psychological Methods*, *20*(4), 523-536. DOI: 10.1037/met0000031

Kwok, O., West, S. G., Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, *42*(3), 557-592. DOI: 10.1080/00273170701540537

Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. *Structural Equation Modeling*, *14*(4), 581–610. DOI: 10.1080/10705510701575438

Liu, Y., Millsap, R. E., West, S. G. Tein, J., Tanaka, R., & Grimm, K. J. (in press). Testing measurement invariance in longitudinal data using ordinal variables. *Psychological Methods*.

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

Meade, A. W., & Lautenschlager, G. K. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*(1), 60–72. DOI: 10.1207/S15328007SEM1101_5

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology*, *95*(4), 728-743. DOI: 10.1037/a0018966

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington, D.C.: American Psychological Association.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York: Routledge.

Millsap, R. E., & Cham, H. (2012). Investigating factorial invariance in longitudinal data. In B. Laursen, T. D. Little, & N.A. Card (Eds.), *Handbook of developmental research methods* (pp. 109-127). New York: Guilford.

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

Murphy, D. L., Beretvas, S. N., & Pituch, K. A. (2011). The effects of autocorrelation on the curve-of-factors growth model. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 430-448. DOI:10.1080/10705511.2011.582399

Muthén, B. O. (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. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in M*plus*. Downloaded from http://www.statmodel.com/download/webnotes/CatMGLong.pdf.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determined power. *Structural Equation Modeling*, *9*(4), 599–620. DOI: 10.1207/S15328007SEM0904_8

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

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: when are statistically significant effects practically important?. *Journal of Applied Psychology*, *89*(3), 497-508. DOI: 10.1037/0021-9010.89.3.497

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

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological methods*, *11*(4), 402-415. DOI: 10.1037/1082-989X.11.4.402

Weiss, R.E. (2005). *Modeling longitudinal data*. New York: Springer.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10-18. DOI: 10.1111/j.1750-8606.2009.00110.x

Wirth, R. J. (2008). *The effects of measurement non-invariance on parameter estimation in latent growth models* (Doctoral dissertation). Retrieved from ProQuest dissertation and theses database. (UMI No. 3331053).

Wirth, R. J., & 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

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory factor analysis of ordinal variables with misspecified models. *Structural Equation Modeling, 17*(3), 392-423. DOI: 10.1080/10705511.2010.489003

APPENDIX A

MATHEMATICAL DEVELOPMENT SUPPORTING THE CONCLUSIONS OF EACH LEVEL OF LONGITUDINAL MEASUREMENT INVARIANCE FOR ORDERED-CATEGORICAL INDICATORS

This appendix is based on the derivations in Liu et al. (in press). It contains the proof of the implications of achieving each level of longitudinal measurement invariance for ordered-categorical indicators. Achieving longitudinal loading invariance implies that changes in the expected *means* of the *continuous latent responses* are fully accounted for by changes in the latent common factors over time. Achieving longitudinal unique factor invariance implies that (a) changes in the not only the expected *means,* but also the expected *variances* and *within-wave covariances* of the *continuous latent responses* are entirely attributable to changes in the latent common factors over time, and more importantly, (b) changes in the expected *means* and *within-wave bivariate probabilities* of the *ordered-categorical indicators* can be fully explained by changes in the latent common factors over time. The derivations below are based on the standard SEM assumption that the common factor scores are uncorrelated with unique factor scores. To account for the longitudinal nature of the design, the common factors are allowed to freely correlate across time, and additionally, each unique factor is allowed to freely correlate with itself, but *not* with other unique factors, at other measurement occasions. Although the present derivations focus on ordered-categorical CFA models with one latent common factor at each measurement occasion, they can be easily generalized to cases with more latent common factors per measurement occasion.

In the ordered-categorical CFA models, the continuous latent responses X^*_{ijt} underlying the observed ordered-categorical responses X_{ijt} are assumed to be multivariate normally distributed (Muthén, 1984), and they are sliced into the ordered-categorical observed responses by a set of threshold parameters ν for each indicator *j* at each measurement occasion *t*:

$$X_{ijt} = c, \text{ if } \nu_{jtc} \leq X^*_{ijt} < \nu_{jt(c+1)}. \quad \text{(A1)}$$

Assuming that $c = 0, 1, \ldots, C$, the response categories of the ordered-categorical indicators, and that $\{\nu_{jt0}, \nu_{jt1}, \ldots, \nu_{jt(C+1)}\}$ are the threshold parameters for the ordered-categorical indicator *j* at measurement occasion *t* with $\nu_{jt0} = -\infty$ and $\nu_{jt(C+1)} = \infty$, the probability of indicator *j* taking on a value *c* can be calculated as

$$\Pr(X_{ijt} = c) = \Pr(\nu_{jtc} \leq X^*_{ijt} < \nu_{jt(c+1)}) = \Pr(X^*_{ijt} < \nu_{jt(c+1)}) - \Pr(X^*_{ijt} < \nu_{jtc}). \quad \text{(A2)}$$

Given that X^*_{ijt} is assumed to follow a normal distribution, one needs to know the mean and variance of that normal distribution to calculate $\Pr(X^*_{ijt} < \nu_{jt(c+1)})$ and $\Pr(X^*_{ijt} < \nu_{jtc})$.

The longitudinal one-factor CFA model for the continuous latent response X^* underlying the measured ordered categorical indicator is

$$X^*_{ijt} = \tau_{jt} + \lambda_{jt}\eta_{it} + u_{ijt}, \quad \text{(A3)}$$

where τ_{jt} is the intercept (typically constrained to zero to allow for the estimation of the threshold parameters), λ_{jt} is the factor loading, η_{it} is the common factor score for person *i* at measurement occasion *t*, and u_{ijt} is the unique factor score for person *i* on indicator *j* at measurement occasion *t*. Following Equation (A3), the expected means of the continuous latent responses X^*_t can be written as

$$E(\mathbf{X}^*_t) = \boldsymbol{\mu}_{X^*_t} = \boldsymbol{\Lambda}_t \kappa_t, \quad \text{(A4)}$$

where $\boldsymbol{\mu}_{X^*_t}$ is a $J \times 1$ vector of the expected means of the continuous latent responses, $\boldsymbol{\Lambda}_t$ is the factor loading vector, and κ_t is the latent common factor mean at measurement occasion *t*. The expected covariance matrix of the continuous latent responses X^*_t can be written as

$$\boldsymbol{\Sigma}_{X_t^* X_t^*} = \boldsymbol{\Lambda}_t \varphi_t \boldsymbol{\Lambda}_t' + \boldsymbol{\Theta}_{tt}, \tag{A5}$$

where φ_t is the latent common factor variance at measurement occasion *t*, and $\boldsymbol{\Theta}_{tt}$ is the unique factor variance-covariance matrix at measurement occasion *t*.

Achieving longitudinal loading invariance means that the factor loading vector $\boldsymbol{\Lambda}_t$ is invariant over time, and as a result, changes over time in $E(\mathbf{X}_t^*)$, the expected means of the continuous latent responses, can be fully explained by changes in κ_t, the latent common factor mean. However, because the threshold parameters that slice the continuous latent responses into the ordered-categorical measured indicators are freely estimated over time (other than the model identification constraints), the expected means of the ordered-categorical measured indicators cannot be entirely attributed to changes in κ_t.

Now focus on only one indicator *j*. Based on Equations (A4) and (A5), the expected mean of the continuous latent response X_{ijt}^* can be written as

$$E\left(X_{ijt}^*\right) = \lambda_{jt} \cdot \kappa_t, \tag{A6}$$

and the expected variance of X_{ijt}^* can be written as

$$Var\left(X_{ijt}^*\right) = \lambda_{jt} \cdot \varphi_t \cdot \lambda_{jt} + \sigma_{jj(t)}^2 = \lambda_{jt}^2 \cdot \varphi_t + \sigma_{jj(t)}^2, \tag{A7}$$

where $\sigma_{jj(t)}^2$ is the unique factor variance for indicator *j* at measurement occasion *t*. Since the continuous latent responses are assumed to follow a normal distribution, given the expected mean and variance of X_{ijt}^* in Equations (A6) and (A7), Equation (A2) can be written as

$$\Pr\left(X_{ijt} = c\right) = \Pr\left(X_{ijt}^* < \nu_{jt(c+1)}\right) - \Pr\left(X_{ijt}^* < \nu_{jtc}\right)$$

$$= \Phi\left(\frac{\nu_{jt(c+1)} - E\left(X^*_{ijt}\right)}{\sqrt{Var\left(X^*_{ijt}\right)}}\right) - \Phi\left(\frac{\nu_{jtc} - E\left(X^*_{ijt}\right)}{\sqrt{Var\left(X^*_{ijt}\right)}}\right)$$

$$= \Phi\left(\frac{\nu_{jt(c+1)} - \lambda_{jt} \cdot \kappa_t}{\sqrt{\lambda^2_{jt} \cdot \varphi_t + \sigma^2_{jj(t)}}}\right) - \Phi\left(\frac{\nu_{jtc} - \lambda_{jt} \cdot \kappa_t}{\sqrt{\lambda^2_{jt} \cdot \varphi_t + \sigma^2_{jj(t)}}}\right), \quad \text{(A8)}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. For an ordered-categorical indicator X_{ijt} with response categories $c = 0, 1, \ldots, C$, the expected mean is

$$E(X_{ijt}) = \sum_{c=0}^{C} c \cdot \Pr(X_{ijt} = c). \quad \text{(A9)}$$

Given Equation (A8), it can be proved that Equation (A9) can be simplified to

$$E\left(X_{ijt}\right) = C - \sum_{c=1}^{C} \Phi\left(\frac{\nu_{jtc} - \lambda_{jt} \cdot \kappa_t}{\sqrt{\lambda^2_{jt} \cdot \varphi_t + \sigma^2_{jj(t)}}}\right). \quad \text{(A10)}$$

When longitudinal threshold invariance is achieved, λ_{jt} and ν_{jtc} are invariant across measurement occasions. Under such circumstances, based on Equation (A10), changes over time in $E\left(X_{ijt}\right)$ will be determined by three things: (1) the latent common factor mean κ_t, (2) the latent common factor variance φ_t, and (3) $\sigma^2_{jj(t)}$, the unique factor variance for indicator *j*. Hence, to attribute mean changes in the ordered-categorical indicators entirely to changes over time in the latent common factor, having invariant factor loadings and invariant thresholds is not sufficient -- the unique factor variance must also be invariant over time.

When longitudinal unique factor invariance is achieved, λ_{jt}, ν_{jtc}, and the elements in $\boldsymbol{\Theta}_{tt}$ are invariant across measurement occasions. Thus for the continuous latent responses, when unique factor invariance holds, based on Equations (A4) and (A5),

changes over time in $E(\mathbf{X}_t^*)$ can be fully explained by changes in κ_t, and changes over time in $\boldsymbol{\Sigma}_{X_t^* X_t^*}$ can be fully explained by changes in φ_t. For the measured ordered-categorical indicators X_{ijt}, when unique factor invariance holds, based on Equation (A10), changes over time in $E(X_{ijt})$ can be fully accounted for by changes in κ_t and φ_t. Moreover, changes in the within-wave bivariate probability of two ordered-categorical indicators taking on certain values can be fully explained by changes in the latent common factors over time. The proof is as follows:

Consider two ordered-categorical indicators, say X_1 and X_2, at measurement occasion *t*. The probability of X_{1t} and X_{2t} taking on values a and b, respectively, can be expressed as

$$\begin{aligned}\Pr(X_{1t} = a, X_{2t} = b) &= \Pr\left(\nu_{1ta} \le X_{1t}^* < \nu_{1t(a+1)}, \nu_{2tb} \le X_{2t}^* < \nu_{2t(b+1)}\right) \\ &= \int_{\nu_{1ta}}^{\nu_{1t(a+1)}} \int_{\nu_{2tb}}^{\nu_{2t(b+1)}} f_{X_{1t}^*, X_{2t}^*}(x_{1t}^*, x_{2t}^*)\, dx_{1t}^* dx_{2t}^*, \end{aligned} \quad \text{(A11)}$$

where $f_{X_{1t}^*, X_{2t}^*}(x_{1t}^*, x_{2t}^*)$ is the joint probability density function for the bivariate normal latent responses X_{1t}^* and X_{2t}^*, and is completely determined by the expected means of X_{1t}^* and X_{2t}^* [$E(X_{1t}^*) = \lambda_{1t} \cdot \kappa_t$, $E(X_{2t}^*) = \lambda_{2t} \cdot \kappa_t$], the expected variances of X_{1t}^* and X_{2t}^* [$Var(X_{1t}^*) = \lambda_{1t}^2 \cdot \varphi_t + \sigma_{11(t)}^2$, $Var(X_{2t}^*) = \lambda_{2t}^2 \cdot \varphi_t + \sigma_{22(t)}^2$], and the correlation between X_{1t}^* and X_{2t}^*, $\rho_{12(t)}$. Based on the tracing rules, the expected covariance between X_{1t}^* and X_{2t}^* can be calculated as

$$COV(X_{1t}^*, X_{2t}^*) = \rho_{12(t)} \cdot \sqrt{Var(X_{1t}^*) \cdot Var(X_{2t}^*)} = \lambda_{1t} \cdot \varphi_t \cdot \lambda_{2t} + \sigma_{12(t)}, \quad \text{(A12)}$$

where $\sigma_{12(t)}$ represents the within-wave unique factor covariance between u_{1t} and u_{2t}, and is equal zero when the unique factors are uncorrelated within-wave. Thus, $\Pr(X_{1t} = a, X_{2t} = b)$ is determined by four sets of model parameters: (1) factor loadings

λ_{1t} and λ_{2t}; (2) threshold parameters ν_{1ta}, $\nu_{1t(a+1)}$, ν_{2tb}, and $\nu_{2t(b+1)}$; (3) unique factor variances $\sigma^2_{11(t)}$ and $\sigma^2_{22(t)}$ and the within-wave unique factor covariance $\sigma_{12(t)}$; and (4) the latent common factor mean κ_t and latent common factor variance φ_t. To attribute changes over time in $\Pr(X_{1t} = a, X_{2t} = b)$ entirely to changes in the latent common factor, the first three sets of model parameters must be invariant across measurement occasions, that is, longitudinal unique factor invariance must hold.

In the configural, loading, threshold and unique factor invariance models, each unique factor is allowed to freely correlate with itself, but *not* with other unique factors, at other measurement occasions. From the derivations above, one can see that the freely estimated lagged unique factor covariances have no influence on the within-wave characteristics of the ordered-categorical indicators or the continuous latent responses. These lagged unique factor covariances only influence the lagged covariances of the same latent responses and the lagged covariances of the same measured indicators across measurement occasions. The proof is as follows:

Consider the ordered-categorical indicator X_{jt} at measurement occasions 1 and 2. The expected lagged covariance between the corresponding latent responses can be written as

$$COV\left(X^*_{j1}, X^*_{j2}\right) = \lambda_{j1} \cdot \varphi_{12} \cdot \lambda_{j2} + \sigma_{jj(12)} = \rho_{jj(12)} \cdot \sqrt{Var\left(X^*_{j1}\right) \cdot Var\left(X^*_{j2}\right)}, \quad \text{(A13)}$$

where φ_{12} is the common factor covariance between measurement occasions 1 and 2, $\sigma_{jj(12)}$ is the lagged unique factor covariance for the unique factor u_{jt} between measurement occasions 1 and 2, and $\rho_{jj(12)}$ is the correlation between X^*_{j1} and X^*_{j2}. Thus, $\rho_{jj(12)}$ can be expressed as

$$\rho_{jj(12)} = \frac{\lambda_{j1} \cdot \varphi_{12} \cdot \lambda_{j2} + \sigma_{jj(12)}}{\sqrt{Var\left(X_{j1}^*\right) \cdot Var\left(X_{j2}^*\right)}}. \quad \text{(A14)}$$

For the measured ordered-categorical indicator X_{jt}, the probability of taking on value a at measurement occasion 1 and value b at measrument occasion 2 can be written as

$$\begin{aligned} \Pr\left(X_{j1} = a, X_{j2} = b\right) &= \Pr\left(\nu_{j1a} \le X_{j1}^* < \nu_{j1(a+1)}, \nu_{j2b} \le X_{j2}^* < \nu_{j2(b+1)}\right) \\ &= \int_{\nu_{j1a}}^{\nu_{j1(a+1)}} \int_{\nu_{j2b}}^{\nu_{j2(b+1)}} f_{X_{j1}^*, X_{j2}^*}(x_{j1}^*, x_{j2}^*)\, dx_{j1}^* dx_{j2}^*, \end{aligned} \quad \text{(A15)}$$

where $f_{X_{j1}^*, X_{j2}^*}(x_{j1}^*, x_{j2}^*)$ is the joint probability density function for the latent response X_{jt}^* at measurement occasions 1 and 2, which is completely determined by the expected means of X_{j1}^* and X_{j2}^* [$E\left(X_{j1}^*\right) = \lambda_{j1} \cdot \kappa_1$, $E\left(X_{j2}^*\right) = \lambda_{j2} \cdot \kappa_2$], the expected variances of X_{j1}^* and X_{j2}^* [$Var\left(X_{j1}^*\right) = \lambda_{j1}^2 \cdot \varphi_1 + \sigma_{jj(1)}^2$, $Var\left(X_{j2}^*\right) = \lambda_{j2}^2 \cdot \varphi_2 + \sigma_{jj(2)}^2$], and the correlation between X_{j1}^* and X_{j2}^*, $\rho_{jj(12)}$. Based on Equations (A14) and (A15), besides the factor loadings, thresholds, and unique factor variances, $\Pr\left(X_{j1} = a, X_{j2} = b\right)$ is also influenced by $\sigma_{jj(12)}$, the lagged unique factor covariance for the unique factor u_{jt} between measurement occasions 1 and 2.