Assessing Measurement Invariance and Latent Mean Differences with Bifactor Multidimensional Data in Structural Equation Modeling

by

Yuning Xu

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

Approved April 2018 by the
Graduate Supervisory Committee:

Samuel Green, Chair
Roy Levy
Marilyn Thompson

ARIZONA STATE UNIVERSITY

August 2018

ABSTRACT

Investigation of measurement invariance (MI) commonly assumes correct specification of dimensionality across multiple groups. Although research shows that violation of the dimensionality assumption can cause bias in model parameter estimation for single-group analyses, little research on this issue has been conducted for multiple-group analyses. This study explored the effects of mismatch in dimensionality between data and analysis models with multiple-group analyses at the population and sample levels. Datasets were generated using a bifactor model with different factor structures and were analyzed with bifactor and single-factor models to assess misspecification effects on assessments of MI and latent mean differences. As baseline models, the bifactor models fit data well and had minimal bias in latent mean estimation. However, the low convergence rates of fitting bifactor models to data with complex structures and small sample sizes caused concern. On the other hand, effects of fitting the misspecified single-factor models on the assessments of MI and latent means differed by the bifactor structures underlying data. For data following one general factor and one group factor affecting a small set of indicators, the effects of ignoring the group factor in analysis models on the tests of MI and latent mean differences were mild. In contrast, for data following one general factor and several group factors, oversimplifications of analysis models can lead to inaccurate conclusions regarding MI assessment and latent mean estimation.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

Measurement invariance (MI) investigates the extent to which the relationships between latent variables and their indicators do not vary across populations or time points. MI is a desired statistical property of a measurement instrument in that it indicates that the same construct(s) are being measured across groups or occasions. The establishment of MI allows the means for groups to be compared on the latent variable(s) or to assess the relationships of latent variable(s) with external measures across groups (Jöreskog & Goldberger, 1975; Sörbom, 1974). If MI is violated, the interpretation of these results is likely to be misleading.

MI is defined with respect to an explicit number of factors with a specific factor structure. Invariant measures with respect to a particular factor may be noninvariant if they are affected by additional factors that are not specified in the model. In practice, these additional factors often represent undesired or unexpected sources of influences on measures and are considered as the leading cause of the lack of MI in multiple group analysis (Kok, 1998; Meredith, 1993). In both structural equation modeling (SEM) and item response theory (IRT), the lack of MI has been informally conceptualized as the presence of unspecified factors on which populations have different latent distributions (e.g., Ackerman, 1992; Camilli, 1992; Jak, Oort & Dolan, 2009; Jeon, Rijmen & Rabe-Hesketh, 2013; Kok, 1998; Roussos & Stout, 1996; Shealy & Stout, 1993).

Knowing that measures are influenced by a wide variety of unexpected sources in practice, the feasibility of applying a unidimensional model to potential multidimensional data has been an important topic of research. For analyses with a single group, the effects

of fitting a unidimensional model to data with multidimensional structures on parameter estimation have been investigated, especially in the IRT literature. For example, a number of studies have found that IRT item parameter estimates are relatively robust to the presence of additional latent variables if there exists one dominant latent variable (Drasgow & Parsons, 1983; Kirisci, Hsu, & Yu, 2001; Reckase, 1979). Procedures and indices were developed to judge whether a dataset is "unidimensional enough" such that parameter estimation is robust to the presence of multidimensionality in data (e.g., Stout, 1993; Zhang & Stout, 1999). In SEM, dimensionality is typically assessed as a part of the overall model fit. In practice, parameter estimation is generally considered as unbiased if the overall model fit is sufficient, although this conclusion is not always warranted (e.g., Reise, Scheines, Widaman, & Haviland, 2013). Two recent studies examined the biasing effects of having unspecified factors that influence subsets of indicators on model parameter estimation within a bifactor modeling framework (Bonifay, Reise, Scheines, & Meijer, 2015; Reise et al., 2013). The results of the studies found that the size of bias in parameter estimates can be correlated with the relative strength of the general factor to the group factors in the bifactor structure, and that this relationship is moderated by the specific bifactor structures underlying one's data.

The assessment of MI, and subsequent analyses on latent variables, requires the correct specification of measurement model. Specifically, both the number of factors and the relationship between the factors and their indicators need to be correctly specified. In SEM, researchers have investigated methods to assess MI and the consequences of violating MI (e.g., Vandenberg & Lance, 2000), but have not directly considered the effects of misspecifying dimensionality in measurement models in their analyses. The

current study explored the effects of dimensionality misspecification on the assessment of MI and on testing latent mean difference between groups. Data were generated with multidimensional structures from two populations; unidimensional analysis models were fit to these data with different between-group equality constraints. MI and between-group latent mean differences were assessed in a multigroup confirmatory factor analysis (CFA) framework.

A bifactor structure was employed to create data with multidimensional structures. With a bifactor model, a general factor underlies all indicators of the factor and one or more group factors underlie subset(s) of the indicators. The group factors reflect additional common variance among clusters of indicators that typically have similar content or have the same context. All factors are specified to be orthogonal to each other. Bifactor models have become increasingly popular in recent research (e.g., Chen, West, Sousa, 2006; Reise, 2012) and, particularly, have been suggested as an effective tool for assessing dimensionality (Morin, Arens, & Marsh, 2015; Reise, Morizot, & Hays, 2007). In this study, multigroup data were generated based on four bifactor structures with different numbers of group factors and different numbers of indicators per group factor. For each of the bifactor structures, factor loadings and factor means of the group factors were varied across groups.

Next a series of single-factor models with different levels of between-group equality constraints were fit to data generated by the bifactor structures to evaluate the effects of misspecifying a multidimensional model as a unidimensional model on the assessment of MI and factor mean differences. The analysis models included models with no cross-group equality constraints, cross-group equality constraints on factor loadings,

and cross-group equality constraints on factor loadings and intercepts. To assess MI, four fit statistics (i.e., χ^2, CFI, RMSEA, and SRMR) were examined for the nested models. Tests of between-group factor mean differences were conducted and estimates of the factor mean differences were examined for each analysis model.

In the next section, I review the statistical definition of MI in SEM and the prototypical steps in assessing MI and latent mean differences, followed by a review of assessing unidimensionality in the context of both single group and multiple group analyses.

Measurement Invariance and Latent Means

Measurement Invariance in the Framework of Confirmatory Factor Analysis

Measurement invariance is the equivalent functioning of a measurement model across different populations. MI holds over all populations defined by a single grouping variable K with respect to a set of latent variables $\mathbf{W}$, if and only if

$$P_k(\mathbf{X} \mid \mathbf{W}) = P(\mathbf{X} \mid \mathbf{W}), \quad (1)$$

where $\mathbf{X}$ is a $p \times 1$ vector of p observed variables, $\mathbf{W}$ is an $m \times 1$ vector of m latent variables, k is the group membership with $k = 1, 2, \ldots, K$, and P denotes the probabilistic function for $\mathbf{X}$ in terms of $\mathbf{W}$. Equation (1) states that the relationship between observed and latent variables does not vary as a function of group membership. Given two individuals with the same level on the latent variables, the probabilities of obtaining a specific response pattern on the observed variables should be the same regardless of their group membership.

When equation (1) does not hold, MI is violated and a lack of MI is said to exist. Violation of MI implies an individual's performance on $\mathbf{X}$ is not only a function of the

latent variables but also a function of group membership. Given the lack of MI, two sources of group differences on observed variables can be confounded: group differences due to population differences on the latent variables and group differences due to inconsistent measurement functions. Different from random errors in measurement, the inaccuracy due to the lack of MI is consistent over replications. Systematic errors will be present in parameter estimation if the measurement model is not appropriately specified. Researchers have found that violation of MI will cause biased estimates of indicator parameters (e.g., Meade & Lautenschlager, 2004) and, subsequently, problematic estimation of latent variables. For example, studies have shown that comparisons of individuals across groups on the latent variables can be biased if MI is violated and not correctly modeled (e.g., Chen, 2008; Meade & Lautenschlager, 2004; Millsap & Kwok, 2004; Whittaker, 2013; Xu & Green, 2015). Also, when using latent variables as predictors in structural models, the lack of MI on the latent predictors can lead to biased estimation of prediction coefficients (e.g., Chen, 2008; Meade & Tonidandel, 2010).

Confirmatory factor analysis is currently the primary factor analytic method for studying MI in SEM. To use CFA to assess MI, it is assumed that investigators have knowledge about the number of underlying factors and the configural pattern of how observed variables represent each factor. In CFA models, a linear relationship between p observed variables and m latent factors in the kth group is specified as in the following equation

$$X_k = \tau_k + \lambda_k \xi_k + \delta_k, \tag{2}$$

where X is a $p \times 1$ vector of observed scores on the p measured indicators, τ is a $p \times 1$ vector of intercepts of observed variables, λ is a $p \times m$ matrix of factor loadings, ξ is a m

× 1 vector of latent factor scores, and δ is a $p \times 1$ vector of unique factor scores. Assuming a standard factor analytic model where E (δ_k) = 0 and the latent factor scores (ξ_k) and unique factor scores (δ_k) are uncorrelated, the model implied covariance matrix Σ_k and the mean vector μ_k for X_k in the *k*th group can be derived as

$$\Sigma_k = \lambda_k \Phi_k \lambda'_k + \Theta_k \quad (3)$$

and

$$\mu_k = \tau_k + \lambda_k \kappa_k, \quad (4)$$

where Φ is a $m \times m$ matrix of factor variance and covariance, and Θ is a $p \times p$ diagonal matrix of variance of unique factors, and κ is a $m \times 1$ vector of factor means in the *k*th group. Within the CFA framework, MI is defined in terms of the extent to which the equivalence of the model parameters τ, λ, and Θ across the *K* groups is tenable.

A traditional taxonomy of MI in CFA defines four hierarchical levels of invariance from liberal to strict: *configural invariance*, *metric invariance*, *scalar invariance*, and *strict invariance* (e.g., Horn & McArdle, 1992; Meredith, 1993; Millsap, 1997; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). *Configural invariance* denotes that the same number of latent variables is represented in each of the groups, and the patterns of zero and non-zero elements in the factor loading matrices are the same across groups. *Metric invariance* (Horn & McArdle, 1992) or *weak factorial invariance* (Widaman & Reise, 1997) requires identical factor loading matrices across all groups (i.e., $\lambda_k = \lambda$). *Scalar invariance* (Steenkamp & Baumgartner, 1998) or *strong invariance* (Meredith, 1993) requires invariant indicator intercepts across groups in addition to invariant factor loading matrices (i.e., $\lambda_k = \lambda$ and $\tau_k = \tau$). *Strict invariance* (Meredith, 1993) states the unique factor variance of indicators is equivalent across all

groups in addition to equivalent factor loadings and intercepts (i.e., $\lambda_k = \lambda$, $\tau_k = \tau$, and $\Theta_k = \Theta$).

For any measure that exhibits some level of invariance but does not demonstrate strict invariance, it is possible that, in any element of λ, τ, and Θ, some indicators are invariant and some are not. *Partial invariance* is defined as the inclusion of both invariant and noninvariant indicators within any defined level of MI except for configural invariance (Byrne et al., 1989; Vandenberg & Lance, 2000). For example, a test can be partially metric invariant, meaning that only a subset of items in the test are equivalent in terms of factor loadings and the rest items have noninvariant factor loadings across groups.

Assessment of MI and Latent Means

Prototypical steps to assess MI. To assess MI, hypotheses are typically tested in a stepwise process. To test configural invariance, a factor model with the same factor loading pattern for all groups is fit to data; no cross-group constraints are imposed except the ones necessary for model identification. For each factor in the model, one indicator is chosen for assigning a metric for the factor, referred to as the referent indicator (RI). Inappropriate selection of RIs is likely to lead to biased estimates of model parameters and inadequate model fit initially before any other decision about invariance is made (e.g., Johnson, Meade, & DuVernet, 2009). Adequate fit of the configurally invariant model indicates that the hypothesized factor structure is supported for all the assessed groups. Failure to retain the model suggests that different underlying factor structure patterns are required for different groups, and group comparisons based on the

hypothesized factor structure are no longer meaningful. As a result, the configurally invariant model is often taken as a baseline model in the assessment of MI.

Given adequate fit for the configurally invariant model, the fit of a metric invariant model with cross-group constraints on factor loadings is compared with the fit of the configurally invariant model. Metric invariance is considered to hold if constraints on factor loadings fail to produce meaningful lack of fit. Similarly, scalar invariance is assessed by comparing the fit of a scalar invariant model with cross-group constraints on both factor loadings and intercepts with the fit of the metric invariant model. Once scalar invariance is achieved, any systematic group differences in the means of observed variables can be attributed to differences in the population means of latent variables. Potentially the next step would be to assess strict invariance using the same strategy. The establishment of strict invariance implies that any systematic difference in the covariance matrices and/or means of observed variables across groups are due to their differences in the latent distributions. Statistically the invariance of unique variance is not a necessary requirement for tests of latent means (Bollen, 1989, pp. 365-369; Byrne et al., 1989; Millsap, 2012, pp. 102 - 109; Vandenberg & Lance, 2000). As a result, strict invariance is not considered in this study. Within any level of invariance except for configural invariance, partial invariance can be tested by comparing models with and without invariance constraints imposed on specific indicators, given the failure to achieve a complete level of invariance. This process of testing MI is discussed in detail in a variety of articles and chapters (e.g., Byrne & Stewart, 2006; Millsap, 2012; Thompson & Green, 2013; Vandenberg & Lance, 2000).

Fit indices to assess lack of MI. Hypotheses about MI can be evaluated by examining how models with different levels of invariance constraints fit the data. Common fit indices include the χ^2 statistic and a variety of goodness-of-fit indices. To compare the relative fit of two nested models (e.g., metric invariant model vs configurally invariant model), a χ^2 difference test is commonly conducted as a formal hypothesis testing method. Simulation studies have found that the χ^2 difference test for assessing group differences in factor loadings adequately controls the Type I error rate and provides relatively high power when used with ML estimator and normally distributed data (French & Finch, 2006). In addition to the χ^2 difference test, one can examine the differences in goodness-of-fit indices from fitting two factor models to evaluate the equivalence of parameters. Previous studies on the sensitivity of fit indices to a lack of MI have found particular fit indices that are sensitive to model misspecifications regarding parameter equality across groups (e.g., Chen, 2007; Cheung & Rensvold, 2002; Fan & Sivo, 2009; Meade, Johnson, and Braddy, 2008). For instance, Cheung and Rensvold (2002) examined 20 goodness-of-fit indices for their changes when cross-group constraints are imposed on factor loadings. Based on a variety of simulation conditions with small to moderate sample sizes, three goodness-of-fit indices (i.e., ΔCFI, ΔGamma, and ΔMcDonald's) were recommended for assessing MI because they are independent of both model complexity and sample size. However, these indices are found to be sensitive to model size for assessing factor mean differences when the mean structures are incorporated (Fan & Sivo, 2009). Chen (2007) examined the sensitivity of five fit indices to a lack of MI at three levels: metric, scalar, and strict invariance. CFI and RMSEA were found to perform equally well to all the three levels of noninvariance, and SRMR

appeared to be more sensitive to a lack of metric invariance than to a lack of scalar and strict invariance. Cutoff values for these indices in assessing MI were recommended in these studies. Across the various conditions in these simulation studies, the proposed cutoff values for examining metric, scalar, and strict invariance ranged from .010 to .021 for RMSEA and -.010 to -.005 for CFI. The cutoff values for SRMR ranged from .005 to .030, depending on the level of MI being examined. Following these findings, this study focused on the χ^2 difference statistic and three goodness-of-fit indices (i.e., RMSEA, CFI, and SRMR) in assessing MI.

Testing latent means and partial invariance. Mean differences of factors can be tested after the establishment of (partial) scalar invariance. Testing factor mean differences across groups is one important application of assessing MI. Compared to other approaches such as multivariate analysis of variance or creating composite scores, the latent variable approach for testing multivariate means minimizes problematic effects of errors in measurement and provides meaningful interpretations of group differences (Cole, Maxwell, Arvey, & Salas, 1993; Hancock, Lawrence, & Nevitt, 2000). To test latent mean differences, two sets of models are specified: one with the factor means constrained to be equivalent across groups (restricted model) and the other with the means freely estimated (full model). The rest of the models are specified as determined through the previous steps of assessing MI. The fit of the two models is compared using a χ^2 difference test. If the increment of fit is significant from the restricted model to the full model, the factor means are considered to be different across groups.

The test of factor mean differences and many other latent variable analyses involving multiple groups require the establishment of MI at a particular level.

Traditionally, opinions state that a full metric invariance should be hold before testing scalar invariance; and only when a full scalar invariance holds, can one proceed to factor mean analysis (Bollen, 1989; Horn & McArdle, 1992). Different factor loadings and intercepts across groups would indicate that individuals with the same factor scores will likely result in different observed scores for the different groups. Alternative views support that partial invariance in terms of factor loadings and intercepts is sufficient for factor mean inferences (Carle, Millsap, & Cole, 2008; Byrne et al., 1989; Marsh & Hocevar, 1985). With partial invariance, only a subset of indicators with invariant factor loadings and intercepts is required for assessing factor mean differences. Statistically, testing of factor means is warranted as long as one of the indictors holds the invariance property. However, to the extent that more indicators are allowed to be different, estimation of factor mean differences are based on a limited number of indicators, resulting in a loss of interpretation in the estimated mean differences, as well as a loss of power in parameter estimation and model parsimony (Green & Thompson, 2012).With a partial invariance assumption, invariance constraints should be imposed on loadings and intercepts that are equivalent across groups in the population, and the remaining loadings and intercepts should be freely estimated in analysis models. If the parameters are improperly constrained across groups, estimates of factor mean differences can be biased. Several studies have found that pseudo-group difference in factor means can appear if cross-group loadings and/or intercepts are falsely constrained to be invariant in analysis models (Chen, 2008; Wang, Whittaker & Beretvas, 2012; Whittaker, 2013; Xu & Green, 2015). As suggested in these studies, bias in estimates of factor mean differences increases as the differences in factor loadings and/or intercepts increases uniformly

between groups. Moreover, incorrectly constraining intercepts has a greater impact on factor mean estimation than incorrectly constraining factor loadings (Chen, 2008; Xu & Green, 2015).

Assessing Dimensionality and Multiple Group Analysis

In assessing MI, an important assumption is that there exists a fixed number of latent variables to characterize the covariance among observed variables. In other words, the establishment of MI relies on a clear definition of dimensionality underlying data. This section discusses the assumptions of model dimensionality and the assessment of unidimensionality in the context of single and multiple group analyses.

Dimensionality and Conditional Independence

Latent variable models generally assume the common variance among a set of variables can be accounted for by a fixed number of underlying factors. If the factor structure is correctly specified, observed variables should be uncorrelated with each other after controlling for factors. This assumption is referred to as conditional independence or local independence, and can be considered as a function of dimensionality (McDonald, 1981). Conditional independence rules out any association among observed variables given the assumed factor structure. With conditional independence, performance on any indicator of a measure should be affected by only individuals' level on the hypothesized latent variables rather than their performance on any other indicators of the measure. Violation of conditional independence indicates that the investigated data does not match the hypothesized dimensionality in a strict sense. The occurrence of conditional dependence can have serious consequences in regard to the applicability of the

hypothesized latent variable models and can lead to biased model estimation as demonstrated in several cases (e.g., Steinberg & Thissen, 1996; Yen, 1993).

Model dimensionality can be explored using exploratory approaches if researchers have little prior knowledge. Exploratory factor analysis is a common method for data reduction and factor structure exploration. The number of factors is determined by synthesizing researchers' substantive knowledge of the dataset and evidence from statistical analyses. Typical analyses for determining the number of factors include the eigenvalue-larger-than-one rule, parallel analysis, and scree plots, which are all based on evaluation of the eigenvalues of correlation matrix of observed variables.

On the other hand, hypothesized factor structures can be tested in a confirmatory way by evaluating how the factor structures fit to empirical data. By conducting CFA, dimensionality is evaluated as an integrated part of the overall model fit (e.g., Swaminathan, Hambleton, & Rogers, 2007). It is assumed that one or more factors are sufficient to characterize data if the factor model fits the data adequately according to fit statistics. When model fit is inadequate, the residual covariance matrix of observed variables is often examined to detect where the misfit might be. Indices based on these residual terms such as SRMR can offer a general view of violation of conditional independence.

Assessing Unidimensionality

Researchers develop measurement instruments with hypothesized dimensions that are sufficiently broad, but at the same time parsimonious, to capture the latent constructs of interest. Unidimensionality is a central assumption for most models within classical test theory and item response theory and has been widely hypothesized in empirical

research (Lord, 1980; McDonald, 1999). Unidimensional measures are desirable in that they are less open to misinterpretation; that is, higher scores on a unidimensional measure can be due only to the single underlying factor rather than some combination of factors. In many multistep modeling procedures, establishing a unidimensional measure is an important preliminary step before conducting the additional required analysis.

In theory, unidimensionality is a plausible assumption when a set of measures is designed to assess a unitary construct. However most measures in practice are unlikely to yield strictly unidimensional data for various reasons. For example, educational assessments measuring achievement and ability levels are often multidimensional. The multidimensionality can rise because a test requires several skills at the same time such as mathematics and reading, both of which have impact on different items in the test to the different extent. Multidimensionality also can emerge due to the multifaceted property of a single broad skill or construct. A mathematical achievement test can be multidimensional as it assesses both the general cognitive ability and the abilities on several specific topics including algebra, geometry, and trigonometry. Similarly, psychological constructs are often characterized by several related facets that are governed by one underlying attribute tendency. Performance on these tests or measures is determined by one's level on both the overall dimension and the dimensions that are content or context specific (e.g., McDonald & Mok, 1995; Reise, Moore, & Haviland, 2010).

Common multidimensional models include correlated factor models and higher-order factor models. The former model assumes a number of correlated latent variables with each accounting for the covariance among a cluster of indicators. The later model

characterizes multidimensionality by constructing one or more second-order factors to account for the covariance among the first-order factors that are often content or context specific (Gustafsson & Balke, 1993). In recent studies, a bifactor structure has become popular to characterize measures that are designed to assess broad constructs (e.g., Gibbons et al., 2008; Morin et al, 2015; Reise et al., 2007; Reise et al., 2010). Bifactor models assume a general factor and a number of group factors. In practice, the general factor usually represents a dominant factor that the test is purported to measure, whereas the group factors are likely to be smaller factors that are context or content specific. Bifactor models are specified such that the group factors are independent of each other and with the general factor. Because of the independence among factors, the variance of indicators in a bifactor model can be separated into three parts: the variance accounted for by the general factor, the variance accounted for by a specific group factor, and the residual variance. The separation of variance in bifactor models is an important advantage in describing factor analytic results (Chen et al., 2006; Reise, 2012).

Researchers have used bifactor models to construct tests and item banks (Gibbons et al., 2008) and applied bifactor models as an alternative approach to other testlet-based models (Chen et al., 2012; DeMars, 2006). Recent substantive research has shown that a bifactor structure is useful in interpreting factor analytic results of measures relative to correlated factor structures and hierarchical factor structures (Reise et al., 2010; Reise, 2012). For example, Reise et al. (2010) demonstrated how to conceptualize an alexithymia scale using a bifactor structure. For this scale, the general factor characterizes the “core” features of alexithymia, and the group factors represent different sub-traits of alexithymia.

Statistically it is desirable to fit multidimensional data with multidimensional models to avoid model misfit and parameter bias. However, unidimensional models frequently are employed to multidimensional data, given the difficulty in defining and interpreting multiple dimensions as well as the complexity in the application of multidimensional models (Kirisci et al., 2001). As a result, assessing whether a unidimensional model can be a sufficient approximation for empirical data becomes an important topic of research (Hattie, 1985; Embretson & Reise, 2000). Researchers argued that one should assess the adequacy of approximate unidimensionality rather than evaluating whether data is strictly unidimensional (Nandakumar & Stout, 1993; Stout, 1987). A unidimensional model can be applicable for multifimensional data if the resulting parameter estimates are relatively unbiased, stable, and consistent.

In the IRT literature, an appreciable body of research has been conducted to investigate the effects of fitting unidimensional models to multidimensional data on estimation accuracy of item parameters and ability distributions (e.g., Ackerman, 1989; De Ayala, 1994; DeMars, 2006; Drasgow & Parsons, 1983; Kirisci et al., 2001; Oshima & Miller, 1990; Reckase, 1979). Although results from these studies are inconclusive because they differed in simulation conditions, analysis methods, and evaluation criteria, a general finding is that the robustness of model estimation to a violation of unidimensionality is closely related to if there exists a strong general factor (Reise, Cook, & Moore, 2015). Unidimensional models are considered as generally applicable for data with one dominant dimension and several minor dimensions. Studies have demonstrated that the estimation of IRT item parameters and latent traits are relatively unbiased if there exists one strong general dimension (e.g., Drasgow & Parsons, 1983; Reckase, 1979). For

example, Reckase (1979) showed that good calibration of items and reasonable ability estimates can be obtained if the first extracted factor accounts for 20% of the variance of a test consisted of 50 items using the 1PL and 3PL models. In the work of Drasgow and Parsons (1983), bifactor data were generated based on factor models with five correlated factors using the Schmid-Leiman transformation method (Schmid & Leiman, 1957). Given the different levels of correlations between the factors, the transformed bifactor data had varying levels of strength for the general and group factors. The results showed that both the estimated item parameters and the latent trait estimates based on a unidimensional IRT model reflected the general factor when the strength of the general factor was moderate or higher. Similarly, for data with several dimensions of approximately equal strength, studies showed the feasibility of applying unidimensionality depends on the pairwise correlations between dimensions (Ackerman, 1989; Kirisci et al., 2001; Oshima & Miller, 1990). The application of unidimensional IRT models is considered as feasible if the dimensions are moderately to highly correlated ($r > .4$), whereas multidimensional models are recommended if the correlations are low and/or vary across dimension pairs (Kirisci et al., 2001).

In addition to examining the effects of violating unidimensional assumptions, methods for assessing the degrees of unidimensionality have been developed in the context of IRT. One common approach to explore "unidimensional enough" is the examination of item covariance residual after fitting a unidimensional model. Derived from the weak form of conditional independence (McDonald, 1981), the DIMTEST can be used to assess the degree of "essential" unidimensionality (Stout, 1987; Nandakumar & Stout, 1993). An essential unidimensionality is hold if, on average, the conditional

covariances of item pairs in a test tend to approach zero as the number of items become larger. Researchers also developed measures such as the DETECT index (Zhang & Stout, 1999) to directly assess the degree of multidimensionality displayed in data, assuming the existence of a dominant single dimension. However, conditional dependence in item pairs will not always result in large residual values; instead, the conditional dependence may lead to distorted estimates of item loadings (Steinberg & Thissen, 1996). Also, the residual-based approach for examining unidimensionality relies on meaningful cutoff values of residuals, which are hard to determine.

Within SEM, procedures that directly assess the "degree" of unidimensionality have not received much attention. The overall model fit remains as the essential rule for evaluating model dimensionality. With satisfactory model fit, bias in parameter estimates in practice is assumed implicitly to be minimal. In this sense, the commonly used fit statistics, such as the χ^2 statistic and goodness-of-fit indices (e.g., CFI), are used to judge the adequacy of unidimensionality. Fit indices are not particularly sensitive for assessing dimensionality because they are designed in general to evaluate departure of data from a hypothesized model rather than specifically to assess dimensionality (Reise et al., 2013). A unidimensional model with good fit based on SEM fit indices can still yield biased item parameter estimates caused by multidimensionality. Recent studies proposed indices such as *explained common variance* (ECV) (Bentler, 2009; Reise et al., 2010; ten Berge & Sočan, 2004) and coefficient omega hierarchical (*omegaH*) (McDonald, 1999; Zinbarg, Revelle, Yovel, & Li, 2005) to denote the strength of the primary factor compared to other orthogonal factors in multifactor models. Particularly, Reise et al. (2013) and Bonifay et al. (2015) studied the performance of three goodness-of-fit indices

(i.e., RMSEA, CFI, and SRMR) and several factor strength indices (i.e., ECV, *omegaH*, and DETECT) in examining the relationship between data departing from a unidimensional structure and the bias in estimates of factor loadings and structural prediction coefficients. Using data generated from bifactor structures, the results found that the degree of parameter estimate bias depends strongly and inversely on ECV, but the effects are moderated by both the number of group factors and the number of indicators. Specifically, the effects are moderated by the percentage of the elements in the data correlation matrix that are uncontaminated by group factors. Given a high percentage of uncontaminated correlations, structural coefficients are found relatively unbiased even when general factor strength is low relative to group factor strength. Also, both CFI and SRMR appear to be related to factor loading and structural coefficient bias, but are not as predictive as ECV. In general, the studies indicated that bifactor structures with a larger number of group factors and a smaller number of indicators per group factor were found to be "closer" to a unidimensional structure in terms of producing less bias in parameter estimates.

Unidimesionality and Multigroup Analysis

The lack of MI is often conceptualized as the differences across populations on one or more unspecified secondary factors or dimensions. However, it is important to note that the presence of unspecified factors is not in itself sufficient for causing noninvariance. The presence of noninvariance depends on the joint distributions of the hypothesized factor and the unspecified factors. With one or more unspecified factors, a necessary requirement for having measurement noninvariance is that the populations must differ in their distributions on the unspecified factor, conditioning on the

hypothesized factor (see Millsap, 2012, pp. 68-71 for a mathematical proof). Conversely, if individuals from different populations have the same latent distributions on the unspecified factors conditioning on the hypothesized factor, the unspecified factors will not introduce noninvariance even if they are not included in the analysis model.

Although the relationship between the lack of MI and the presence of secondary factors has been considered in the literature, the direct impact of ignoring such factors on the analyses for multiple groups has not been investigated. Violating unidimensionality in multigroup analysis can cause biasing effect of parameter estimation in multiple group analysis. Although all items of a test may be good measures of a hypothesized latent variable across all populations, some of the items might be influenced by additional factors in one or more groups. Different from the single group analysis where the primary focus is on parameter estimate accuracy, the research goals in multiple group analysis are predominantly on comparisons of multiple populations in terms of measurement parameters and latent distributions of the populations. Thus, instead of examining the direct relationship between the presence of secondary factors and parameter estimation bias in each group, the bias in differences in parameter estimates across multiple populations is of more concern; such bias in differences can form errors in judgments about MI and comparisons in the means of latent variables.

CHAPTER 2

STUDY OBJECTIVES

Methodological studies on multiple group analysis have focused on assessing scalar and metric invariance and the consequences of violating these particular levels of MI on parameter estimation (e.g., Byrne, Shavelson, & Muthen, 1989; Chen, 2008; Kaplan & George, 1995; Whittaker, 2013; Yoon & Millsap, 2007). In these studies, the assumption has been made that the same factor structure holds for all investigated groups. The first objective of the current study was to examine the impact of having unspecified secondary factors on the assessment of MI and on tests of latent mean differences. Data were generated using bifactor models and fit to a set of bifactor models and a set of single-factor models with different levels of invariance constraints. Assessments of MI and tests of latent mean differences were conducted for both the bifactor models and the single-factor models following the prototypical steps described earlier. The χ^2 statistics and three goodness-of-fit indices (i.e., RMSEA, CFI, and SRMR) were used to assess a lack of MI. Bias in the estimates of latent mean differences were examined and compared when fitting the bifactor and single-factor analysis models. In addition, a standardized effect size measure of the estimated factor mean differences was computed.

The second objective of the study was to investigate if the results of assessing MI based on fit indices and likelihood ratio tests are informative in indicating the size of bias in estimates of latent mean differences when the dimensionality of the analysis model is misspecified. It was expected that when bifactor data are analyzed with single-factor models, any between-group difference in parameters associated with group factors will be mapped into differences in indicator parameters in the single-factor model; the result will

be a lack of MI. Such measurement noninvariance is the cause of bias in latent mean differences if the analysis model is not respecified.

The study used a series of bifactor structures to characterize data that are primarily unidimensional, but also are affected by one or more secondary factors. Several features of the bifactor models for data generation were varied. First, with a fixed number of indicators, the number of group factors and the number of indicators per factor were manipulated. The purpose was to create varying proportions of nonequivalent parameters of indicators in the single-factor analysis models. Both partial and complete noninvariance at a particular level of MI were created with this manipulation. Second, for any specific bifactor structure, factor loadings and latent means of the group factors were generated to have different values across populations. Population differences in group factor loadings and group factor means were expected to lead to nonequivalent factor loading and intercept estimates, respectively, when fitting a single-factor model. Additionally, varying the magnitude of differences of the group factor parameters was expected to show how these differences translated into the lack of MI, and potentially into the bias of estimating factor mean differences when specifying a single-factor analysis model.

CHAPTER 3

METHODS

The primary purpose of the study was to investigate the impact of unspecified secondary factors on the tests of MI as well as on the test of factor mean differences. In all analyses, data were generated based on bifactor models with nine indicators that loaded on a general factor and subset(s) of the indicators that loaded on one or more group factors for two populations. The latent distributions of the group factors and the factor loading strength of indicators on the group factors were manipulated within and between the populations. In conducting the analyses, a set of bifactor models and a set of single-factor models with varying levels of invariance constraints were fit to the generated data to test for MI and to estimate factor mean differences between the populations. The bifactor analysis models were considered as baseline models for comparisons with the single-factor models. Fit indices and the standardized estimates of factor mean differences were analyzed at each step of the MI tests. These statistics were expected to be diagnostic in detecting measurement model misspecification as increasingly strict invariance constraints were imposed on analysis models.

The study was conducted in two steps. In Study 1, data simulation and analyses were conducted at a population level. Based on the results of Study 1, Monte Carlo simulations were conducted at a sample level for a subset of conditions in Study 1.

Study 1

Simulation conditions. In the generation of the data, four simulation factors were manipulated: the generation factor structure, indicator loadings on the general factor,

indicator loadings on the group factors, and the differences in latent means between populations on the group factors. Details of these manipulations are described below.

Factor structure. Data were generated based on four different bifactor structures for two populations. All latent structures included a general factor associated with nine indicators and one or more group factors associated with subset(s) of the indicators. Figure 1 presents the path diagrams for the four bifactor structures. Structures 1 and 2 were two-factor structures with all indicators loading on a general factor and a subset of indicators loading on a group factor. In structure 1, three of the nine indicators loaded on one group factor; in structure 2, six of the nine indicators loaded on one group factor. Structures 3 and 4 were bifactor structures with all indicators loading on a general factor as well as on one of the multiple group factors. Structure 3 had the first three indicators and the last six indicators loading on two group factors respectively. For structure 4, the first three indicators, the second three indicators, and the last three indicators loaded on each of the three group factors. For all factor structures, group factors were uncorrelated with the general factor and were uncorrelated with each other. In any generation condition, the two populations had the same factor structure.

The four bifactor structures differed in the number of group factors, the number of indicators per group factor, and/or the number of indicators associated with a group factor. Among the four structures, structures 3 and 4 followed a typical bifactor structure where all indicators loaded on a general factor and on one of the group factors. Structures 1 and 2 were less typical in that only a subset of the indicators were associated with a group factor.

General and group factor loadings. A preliminary study was conducted to determine the magnitudes of the general and group factor loadings for Study 1. In testing MI, the typical first step is to assess configural invariance. The same analysis model needs to fit adequately to each population before one can proceed to the next steps. In the current study, when the analysis model is a single-factor model, the first step of analysis is to assess if the single-factor model fits adequately to both populations. The preliminary study was conducted at the population level to explore the effects of the combinations of different magnitudes of general and group factor loadings on the fit of a single-factor model fitting to bifactor data. The fit of the single-factor models was assessed based on three fit indices – CFI, RMSEA, and SRMR. In the preliminary study, covariance matrices for a single population were generated based on different general and group factor loadings for the four factor structures. Standardized general factor loadings were varied at .4, .5, .6, .7, and .8; and standardized group factor loadings were varied at .2, .3, .4, and .5. Covariance matrices were simulated based on all combinations of these magnitudes and were fit using a single-factor model. Tables A1-A4 in Appendix A present the fit results for fitting a single-factor model to covariance matrices generated based on the four bifactor structures with all combinations of the magnitudes for the general and the group factor loadings. Based on the results, .5 and .7 were selected for general factor loadings, and .2 and .4 were selected for group factor loadings. At the population level, these loading values ensured that the fit indices from fitting a single-factor model to data indicated somewhat below adequate to adequate model fit according to the conventional cutoff criteria (e.g., .08 for RMSEA, .95 for CFI, and .08 for SRMR; Hu & Bentler, 1999); therefore, configural invariance could be established across

populations. With the selected factor loadings, the single-factor model fit worst to data with general factor loadings of .5 and group factor loadings of .4 under factor structure 4. For this condition, RMSEA was .10, CFI was .84, and SRMR was .06 (see Table A4).

General factor loading strength. Based on the preliminary study, the loadings on the general factor averaged around .5 or .7. The actual general factor loadings varied around these values. Specifically, for the average loading of .7, the actual general factor loadings were set at .6, .7, and .8 in sets across the nine indicators for a factor structure. Similarly for the average loading of .5, the actual general factor loadings were .4, .5, and .6 in sets across the nine indicators. For all generation conditions, the general factor loadings were kept invariant across groups.

Group factor loading strength. Loadings on the group factors were either invariant or noninvariant across the two populations. The invariant conditions had average group factor loadings of .3 (specified values of .2, .3, and .4) for both populations. For noninvariant conditions, two thirds of the indicators associated with a group factor had different group factor loadings between populations. For these indicators, their noninvariant group factor loadings were .2 for population 1 and .4 for population 2. Table 1 presents the general and group factor loadings for conditions with an average general factor loading of .7. In these noninvariant conditions, population 2 had uniformly greater group factor loadings than population 1 on the specified indicators.

The study also considered noninvariant conditions where two thirds of the indicator loadings on a group factor had different group factor loadings; however, the two populations had equivalent group factor loading values on average. For these noninvariant conditions with non-uniform differences, population 1 had loadings of [.2,

.3, .4] or [.2, .3, .4, .2, .3, .4] for the 3-indicator group factors or the 6-indicator group factors, respectively; in population 2, the group factor loadings were [.4, .3, .2] or [.4, .3, .2, .4, .3, .2]. The nonuniform conditions were examined for only a subset of the conditions of the simulation design, that is, with general factor loadings of .7 and group factor mean differences of .4.

For both the general and group factors, the variation in the actual loadings around the average magnitude was purposely designed to avoid empirical under-identification of the bifactor analysis models found in the pilot study. Specifically, when fitting a bifactor model to data generated from a bifactor model with uniform general and group factor loadings (e.g., all general factor loadings equal to .7 and all group factor loadings equal to .2), the mean structure of the model was empirically under identified. Varying the factor loadings slightly across the indicators resolved the identification issue. See Green and Yang (2017) for further discussion of this issue.

Group factor mean differences. Across all generation conditions, the latent scores on the general factor followed a normal distribution with a mean of 0 and a variance of 1 in both populations. The latent scores on the group factor(s) were normally distributed with variances of 1, but with either the same or different means in the two populations. In population 1, the group factor mean(s) were set at 0 for all generation conditions. In population 2, the group factor mean(s) were varied: 0, .2, and .4. For factor structures with more than one group factor, the means of all group factors within a population were kept the same. The values of .2 and .4 reflected small and small-to-medium effect sizes of latent mean differences (Hancock, 2001) and were expected to

result in a lack of scalar invariance as well as bias in estimates of factor mean differences when analysis models were misspecified as unidimensional.

For all generation conditions, the intercepts of all indicators were 0 for both populations. The residual variances were set at one minus the variance accounted for by the general factor and the group factors.

The simulation design yielded a total of 52 conditions. Data covariance matrices were generated and were analyzed at the population level.

Analysis models and assessment criteria. A set of bifactor models and a set of single-factor models were specified as analysis models. These included bifactor and single-factor models with no between-group equality constraint (configurally invariant models), with between-group equality constraints on factor loadings (metric invariant models), and with between-group equality constraints on both factor loadings and intercepts (scalar invariant models). A very large sample size of 1,000,000 was used for all model fitting to mimic analyses at the population level.

In fitting the bifactor analysis models, one indicator for each group factor that had invariant loading across populations was selected as the RI for the factor. For the general factor in each factor structure, all indicators were invariant in terms of the loadings and intercepts so RIs were chosen arbitrarily: indicator 1 in analysis models that were consistent with structures 1, 3, and 4, and indicator 4 in analysis models that were consistent factor structure 2. For the RIs, factor loadings were fixed at 1 in both populations, and the intercepts were constrained to be equivalent across populations. Variances of all factors were allowed to be freely estimated for both populations. All factor means were fixed at 0 for population 1 and were freely estimated for population 2.

The selection of the RI for fitting the single-factor analysis models depended on the specific factor structures. For structures 1 and 2, the RIs were arbitrarily chosen from the indicators that were not associated with a group factor so that the RIs were truly invariant in the single-factor model (i.e., indicator 5 in Figure 1a and indicator 2 in Figure 1b). In structure 3, indicator 2 (see Figure 1c) was selected as the RI because it was invariant in terms of both the general and group factors. In structure 4, indicator 2 was arbitrarily selected as the RI as shown in Figure 1d because it was invariant on both the general and group factors. For all single-factor analysis models, the factor means were fixed at 0 for population 1 and were freely estimated for population 2. Factor variances were freely estimated for both populations.

For each of the analysis models, changes in fit indices, including RMSEA, SRMR, and a revised CFI, were examined for the nested invariant models. The revised CFI was computed using an appropriate baseline model (i.e., a baseline model nested within the analysis models) as suggested by Widaman and Thompson (2003) for MI assessment. Revised CFI will be referred to as simply CFI in the remainder of the manuscript. Estimates of mean differences on the general factor of the bifactor models and on the single factor of the single-factor models were examined. Bias in the estimates and standardized effect size statistics for the estimates were computed. The standardized effect size statistics were calculated using the estimated factor mean differences and the pooled variances from the two populations.

All bifactor models were correctly specified in terms of dimensionality, but the models were misspecified in the assessment of metric invariance when data were generated to have noninvaraint group factor loadings. All single-factor models were

misspecified in terms of dimensionality. As a result, the metric invariant analysis models were expected to exhibit a lack of fit when data were generated with noninvariant group factor loadings. The scalar invariant models were expected to exhibit a lack of fit when group factor means were generated to be different between populations.

Study 2

A subset of the simulation conditions in Study 1 at the population level were conducted in Study 2 at the sample level. Specifically, Study 2 included conditions where the between-group group factor means differences were .4. Datasets were generated given the four different bifactor structures, two levels of general factor loadings (.7 or .5), and invariant or noninvariant group factor loadings. Study 2 also investigated two levels of sample sizes: 150 or 300 in each group. Factors and errors were generated to be normally distributed. The same set of analysis models investigated in Study 1 were applied in fitting sample data in Study 2. The design yielded a total of 32 simulation conditions. For each simulation condition, 1000 replications of sample datasets were generated and analyzed.

Analysis models and assessment criteria. As in Study 1, the sets of bifactor and single-factor analysis models were fit to sample data. For each of these models, fit indices including CFI, RMSEA, and SRMR and the estimated factor mean differences were analyzed for their replication means. In addition, the χ^2 difference tests were conducted for nested invariant models to assess different MI levels. For the bifactor analysis models, the empirical rejection rates for the Wald test of the general factor mean differences were examined at the .05 level. For the single-factor analysis models, the empirical rejection

rates were examined for the Wald test of the mean differences on the single-factor at the .05 level.

For both Studies 1 and 2, the simulation and analysis work were conducted using R 3.1 and *M*plus 6.11.

CHAPTER 4

RESULTS

Results of Studies 1 and 2 (i.e., results at the population and sample levels) are summarized in each of three sections. The first section summarizes model convergence for bifactor analysis models. In the second section, the effects of fitting bifactor and single-factor models to bifactor data on assessing MI at particular levels were examined. The third section examined the bias and the standardized effect size statistic for the estimates of between-group factor mean differences by fitting the bifactor and single-factor analysis models.

Model Convergence

Fitting bifactor models to bifactor data resulted in out-of-bound parameter estimates for generation conditions under structure 3 at the population level. At the sample level, the bifactor analysis models resulted in different numbers of replications that did not converge and/or had model solutions with out-of-bound parameter estimates across all generation conditions. In contrast, fitting single-factor models to bifactor data never resulted in any improper model solution at the population or the sample level.

Improper solutions when fitting bifactor models at the population level appeared for all generation conditions under structure 3 (the two-group-factor structure) if the model was misspcified. The estimates for the variances of one of the two group factors in group 1 were found to be negative when the group factor loadings of the bifactor analysis model were incorrectly constrained to be invariant. In addition, fitting bifactor metric invariant models to generated data with non-uniform noninvariant group factor loadings also led to improper solutions. The improper solutions were observed for three of the four

factor structures but not for structure 2. Under all the other generation conditions, solutions based on fitting a bifactor model converged properly when the invariance of group factor loadings were misspecified.

At the sample level, fitting bifactor analysis models to the data generated for the 1000 replications for each generation condition resulted in a limited number of replications that reached proper solution (i.e., successful model convergence with no out-of-bound parameter estimates). Figure 2a and Figure 2b present the numbers of replications that reached proper solution across the three analysis models (i.e., configurally invariant model, metric invariant model, and scalar invariant model) for each generation condition. Figure 2a and 2b present the results for conditions with general factor loadings of .7 and .5, respectively. Only replications with proper solutions were included in the analyses at the sample level.

Figures 2a and 2b evidence the same patterns across and within the generation factor structures. Conditions with higher loadings on the general factor had higher convergence rates. Across the structures, conditions with only one group factor (structure 1 and structure 2) had better convergence rates than conditions with multiple group factors (structure 3 and structure 4). Comparing the two one-group-factor structures, having a group factor with more indicators (structure 2) had better convergence rates. For the multiple-group-factor structures, having group factors with equal numbers of indicators (structure 4) resulted in higher convergence rates than having group factors with unequal numbers of indicators. Within each factor structure, conditions with larger sample size had higher convergence rates as expected. The effects of the invariance of group factor loadings on convergence rates depends on the specific factor structures.

Under structures 1, 3, and 4, having noninvariant group factor loadings generally led to better convergence rates. While under structure 2, the conditions with invariant group factor loadings had higher convergence rates.

Within each generation condition (not shown in Figures 2a and 2b), imposing invariance constraints on analysis models led to increased convergence rates, regardless of whether or not the parameters in the generation models were invariant. This is not surprising because imposing constraints to analysis models leads to greater numbers of degrees of freedom. The numbers of replications that reached proper solutions for all analysis models with different levels of invariance constraints are presented in Tables B1 and B2 in Appendix B.

Assessing Measurement Invariance

Assessing configural invariance. Configural invariance was assessed by fitting bifactor and single-factor models with no invariance constraints imposed on the between-group parameters (except for model identification purpose) to bifactor data at the population and the sample levels. As expected, bifactor models fit perfectly to data at the population level and had excellent fit at the sample level. The single-factor models fit worse than bifactor models at both population and sample level. The degree of misfit for single-factor models differed as a function of the generation factor structures.

Assessment of configural invariance for fitting bifactor analysis models. Bifactor models fit perfectly to bifactor data for all generation conditions at the population level. At the sample level, bifactor models fit almost perfectly to the data for all generation conditions; the average RMSEAs were smaller than .03, average CFIs were greater than .98, and average SRMRs were smaller than .05. Figures 3a and 3b present

the averages of the χ^2 statistic and the three fit indices for conditions with group factor mean differences of .4. Results for conditions with group factor mean differences of 0 and .2 are not shown because the results differed from those for the .4 conditions only in the fourth decimal place. Figure 3a presents the results for conditions with general factor loadings of .7, and Figure 3b presents the results for conditions with general factor loadings of .5. Figures 3a and 3b demonstrated similar patterns for the four fit statistics. In both figures, conditions under structure 3 had either equal or smaller average χ^2 than the corresponding conditions under structure 4 with the same degrees of freedom. The average RMSEA, CFI, and SRMR indicated slightly better fit for structures 2 and 3 than for structures 1 and 4. The former two structures both contained a group factor with six indicators. Increasing sample size from 150 to 300 led to better SRMR across all generation structures, but only for conditions under structures 1 and 4 for RMSEA and CFI. Comparing the Figure 3a and Figure 3b, conditions with weaker general factor loadings (.5) had slightly worse fit based on the fit indices, particularly for conditions under structures 1 and 2.

Assessment of configural invariance for fitting single-factor models. Fit of single-factor models to bifactor data depended highly on the generation factor structures. Figure 4 presents the fit indices for fitting single-factor models at the population level for conditions with group factor mean differences of .4. Figures 5a and 5b present the averaged fit indices at the sample level for conditions with group factor mean differences of .4. Additionally, the effect of uniform and non-uniform noninvariant group factor loadings was also examined. The results are presented in Figure C1 in Appendix C for conditions with general factor loadings of .7 and group factor mean differences of .4 at

the population level. Results for conditions with group factor mean differences of 0 and .2 are not shown because the results differed from the .4 conditions only in the fourth decimal place.

A similar pattern of results was demonstrated at the population and sample levels except that the average SRMRs at the sample level were higher than SRMRs at the population level for the corresponding conditions (i.e., Figures 4 and 5, respectively). When fitting single-factor models to data with bifactor structures with only one group factor (structures 1 and 2), the fit was adequate, but was slightly worse than the fit for bifactor analysis models. Compared to structure 1, structure 2 where more indicators were associated the group factor had better fit based on all the indices. When the generation factor structures had more than one group factors (structures 3 and 4), fit for the single-factor analysis models became substantially worse than fit for the bifactor analysis models. The single-factor model fit worst for the three-group-factor structure (structure 4). For conditions with generation models under structure 4, the RMSEAs were all greater than .06 and the CFIs were all below .95 at the population level.

Conditions with uniform and non-uniform, noninvariant group factor loadings had comparable fit for fitting single-factor configurally invariant models at the population level, as shown in Figure C1. All three fit indices agree that the non-uniform conditions had slightly better fit than the uniform conditions for all structures except structure 2.

Across the factor structures, CFIs indicated better fit for the single-factor models when the data were generated with stronger general factor loadings of .7, whereas χ^2 and RMSEAs yielded results with the opposite interpretation. The latter pattern requires further investigation. This pattern is observed at both the population level (shown in

panels (b) and (a) of Figure 4) and the sample level (shown in panel (c) of Figures 5a and 5b for CFI, panel (a) of Figures 5a and 5b for chi-square, and panel (b) of Figure 5a and 5b for RMSEA). At the sample level, increasing sample size from 150 to 300 led to larger chi-squares and SRMRs (shown in panels (a) and (d) of Figure 5).

Assessing metric invariance. To assess metric invariance, bifactor and single-factor models with invariance constraints imposed on all between-group factor loadings were fit to bifactor data at the population and the sample levels. Fit for the metric invariant models was then compared to fit for the configurally invariant models. The presentation of the results of assessing metric invariance is divided into two parts. The first part discusses the results for generation conditions with group factor loadings generated to be invariant, and the second part discusses the results for conditions with noninvariant group factor loadings.

Assessment of metric invariance for generation conditions with invariant group factor loadings. At the population level, for generation conditions with invariant group factor loadings, imposing invariance constraints on factor loadings led to no change in CFIs and SRMRs, and improved fit based on RMSEAs, regardless of whether the data were analyzed with a bifactor or a single-factor models. The improvement in RMSEA from the configurally invariant model to the metric invariant model demonstrated the penalty for model complexity of the index. With no additional misspecification, more parsimonious models are preferred based on RMSEA.

At the sample level, the increases in average RMSEAs by imposing invariance constraints on factor loadings were smaller than .001 across conditions, regardless of the

analysis models. The decreases in average CFIs were smaller than .006, and the increases in average SRMRs were smaller than .02.

Also at the sample level, the empirical rates of rejecting a metric invariant model for each generation condition were examined when fitting bifactor and single-factor models. The rejection rates were deemed as empirical Type I error rates when fitting bifactor models and pseudo Type I error rates when fitting single-factor models (because the single-factor models are misspecified models in terms of dimensionality). In Figure 6, I present the Type I and pseudo Type I error rates when assessing metric invariance for conditions with group factor mean differences of .4. Figure 6 panel (a) shows that Type I error rates for assessing metric invariance for the bifactor analysis models fell within the acceptable range of .025 to .075 (Bradley, 1978), with four exceptions. The exceptions occurred when sample size was 150 and when sample size was 300 with general factor loadings of .5. All four conditions were generated based on factor structures 3 and 4. For these four conditions, the models failed to properly converge across a large number of replications. The largest number of replications that converged properly was 99 out of the 1000 replications. On the other hand, the pseudo Type I error rates for the single-factor models were comparable to the rates for the bifactor models and had less variability, as shown in panel (b) of Figure 6. No rate for the single-factor model fell outside the range of .025 to .075.

Assessment of metric invariance for generation conditions with noninvariant group factor loadings. For generation conditions with noninvariant group factors, imposing invariance constraints on factor loadings led to worse fit for both bifactor and single-factor models. Table 2 summarizes the changes in fit indices at the population

level for the bifactor and the single-factor models. The changes in fit indices were calculated by subtracting fit values for a configurally invariant model from fit values for a metric invariant model. Underlined values in the table indicated that a metric invariant model fit better than a configurally invariant model based on the specific fit indices, despite the misspecified invariant constraints. Table C1 in Appendix C presents the results for comparing the uniform and non-uniform group factor loading differences when assessing metric invariance with single-factor models.

In Table 2, across generation conditions, CFI was more sensitive to the lack of metric invariance for single-factor analysis models, whereas RMSEA was more sensitive to a lack of metric invariance when bifactor models were the analysis models given the greater change in degrees of freedom. Changes in SRMRs were not consistently greater for either the bifactor or the single-factor analysis models.

As shown in Table C2, uniformity versus non-uniformity of group factor loadings did not have a consistent effect on the sensitivity of fit indices when assessing metric noninvariance with single-factor models. The fit indices had greater sensitivity for the uniform conditions under structure 1, but for the-nonuniform conditions under structure 4.

At the sample level, the empirical rates of rejecting a metric invariant model were examined for the bifactor and the single-factor analysis models. The rejection rates were deemed as empirical power rates for the bifactor models and pseudo power rates for the single-factor models. Figure 7 presents the power and pseudo power rates for assessing metric noninvaraince for the bifactor and single-factor analysis models for conditions with group factor mean differences of .4.

In panel (a) of Figure 7, the highest power of .22 was for the condition under structure 3 with general factor loadings of .7 and sample size of 300. Structure 2 had higher power rates than structure 1 across the other two generation factors; whereas power rates for structures 3 and 4 did not have a uniform pattern. This may due to the small number of replications that converged properly for these structures. Power rates for the single-factor analysis models are presented in panel (b) of Figure 7. Across the different general factor loadings and sample sizes, power rates were higher for factor structures with more indicators per group factor (structures 2 and 3) than for structures with fewer indicators per group factor (structures 1 and 4). Given the same magnitude of differences in group factor loadings for any noninvariant indicator in the generation models, the estimated loading differences when fitting a single-factor model to bifactor data depended on the total number of noninvariant indicators and the number of noninvariant indicators per group factor in the generation structure. The estimated loading differences between groups were found greater for structures 2 and 4, which had more indicators per group factor. The greater differences in loadings lead to higher power rates for rejecting metric invariance. Although structures 3 and 4 had the same number of noninvariant group factors, power rates for structure 4 were lower because of the smaller differences in estimated loadings. For any factor structure, a greater sample size and/or stronger general factor loadings led to higher power rates.

Across all the generation and analysis conditions, the power of rejecting metric invariance was low. The highest power was .25 for fitting single-factor models to bifactor data generated under structure 2 with general factor loadings of .7 and a sample size of 300. Unexpectedly, power rates for the single-factor analysis models were consistently

higher than power rates for the bifactor analysis models across all but two generation conditions. The differences in the power rates between the bifactor and single-factor analysis models may be a function of the numbers of replications that converged properly.

Assessing scalar invariance. Scalar invariance was assessed for generation conditions with invariant group factor loadings. To assess scalar invariance, bifactor and single-factor models with invariance constraints on all between-group factor loadings and intercepts were fit to bifactor data at the population and the sample levels. Fit statistics for the scalar invariant models were compared with fit statistics for the metric invariant models.

Assessment of scalar invariance when fitting bifactor analysis models for generation conditions with invariant group factor loadings. Fitting bifactor scalar invariant models to data generated conditions with invariant group factor loadings led to no change in fit at the population level and little change in fit at the sample level as expected. Panel (a) of Figure 8 presents the empirical rejection rates (Type I error rates) for assessing scalar invariance when bifactor models were analyzed for conditions with group factor mean differences of .4. Across the two sample sizes and the two levels of general factor loadings, Type I error rates for structures 1 and 2 and all but one conditions under structure 4 were within the range of .025 - .075. Four of the five outliers were for conditions under structure 3 and one outlier was for one condition under structure 4. For these conditions, the numbers of replications that reached proper solution were very few. The highest number of replications for conditions under structure 3 was 17; the number of replication for the structure 4 condition was 99.

Assessment of scalar invariance when fitting single-factor analysis models for generation conditions with invariant group factor loadings. Imposing invariance constraints on indicator intercepts led to decreasing model fit when fitting single-factor models to bifactor data generated with *nonzero* group factor mean differences. Such results were consistent with expectation because the single-factor models misspecified data dimensionality.

Table 3 presents changes in fit indices for single-factor models for generation conditions with group factor mean differences of .2 and .4 at the population level. In Table 3, with group factor mean differences of .2 or .4, a negative relationship was observed between the number of indicators associating with a group factor in the generation bifactor structure and the sensitivity of fit indices to a lack of scalar invariance. Specifically, all three fit indices had the greatest changes when intercept constraints were imposed for generation conditions under structure 1 (i.e., only three indicators loaded on a group factor). In contrast, changes in fit indices were minimal for structures with all nine indicators loading on group factors.

Panel (b) of Figure 8 presents the empirical rates of rejecting a scalar invariant model for generation conditions with group factor mean differences of .4. This empirical rejection rate was deemed as pseudo power rate for rejecting scalar invariance for the single-factor models. Similar to results at the population level, the power rates were found to be highest for structure 1 (the structure with the fewest number of indicators loading on one group factor), and decreased as more indicators loaded on group factors in the generation structures across the sample sizes and the two levels of general factor loadings. Similar with assessing metric invariance, a greater sample size and/or stronger

general factor loadings led to higher power rates for detecting scalar noninvariance for a factor structure.

Results in Table 3 and Figure 8 panel (b) are counterintuitive. With nonzero group factor mean differences in the generation factor structures, fit indices were less sensitive to scalar noninvariance when more indicators loaded on group factors. Fitting a single-factor analysis model to bifactor data ignores the group factors. One would expect the mean differences in the ignored group factors are manifested as differences in intercepts of the associated indicators when a single-factor model is analyzed with data. Thus, imposing constraints on the intercepts should lead to worse model fit. This effect was observed under structures 1 and 2 when a subset of indicators loaded on a group factor. For structures 3 and 4, all indicators loaded on group factors with between-group mean differences. Thus, the observed means of all indicators in group 2 were homogeneously higher than the observed means of indicators in group 1. When fitting single-factor models, the homogeneous mean differences of all indicators were manifested as difference in means of the single factor, instead of differences in the intercepts of indicators. Thus, constraining intercepts to be invariant in structures 3 and 4 led to only minor changes in fit indices.

Testing Between-Group Mean Differences

Between-group factor mean differences were examined for generation conditions with invariant group factor loadings. For the bifactor analysis models, I tested the differences in the general factor means between groups. For single-factor analysis models, I tested the differences in the single factor means. With both analysis models,

invariance constraints were imposed on all between-group factor loadings and intercepts, regardless of whether scalar invariance was achieved in the previous step.

Two statistics for the estimates of factor mean differences at the population and sample levels were computed: bias in the estimates and a standardized effect size measure for factor mean differences. At the sample level, the rejection rates of testing the equivalence of the factor means between groups were also summarized.

Testing between-group factor mean differences for bifactor analysis models. For generation conditions with no group factor mean differences, no bias was observed for bifactor and single-factor models at the population level. Table 4 (left half) presents bias and the standardized effect size for bifactor analysis models at the population level for generation conditions with nonzero group factor mean differences (i.e., $\Delta\kappa_{GRP} = .2$ and .4). In Table 4, the bias and standardized effect sizes for bifactor models were all zero except for three conditions with structures 3 and 4 generation models. Panel (a) of Figure 9 presents the empirical Type I error rates for testing equivalent factor means for conditions with group factor mean differences of .4. All Type I error rates were within the acceptable range for conditions under structures 1 and 2. Type I error rates varied from .03 to .41 for conditions under structures 3 and 4 with very few numbers of replications that reached proper solutions.

Testing between-group factor mean differences for single-factor analysis models. The right half of Table 4 presents bias and the standardized effect size for single-factor models at the population level for generation conditions with nonzero group factor mean differences (i.e., $\Delta\kappa_{GRP} = .2$ and .4). Compared to the bifactor analysis models, the bias and standardized effect sizes for single-factor models were much greater. For the

single-factor models, the magnitude of bias and standardized effect sizes varied as a function of the bifactor structures underlying data. Bias and the standardized effect size were greater when more indicators loaded on a group factor in the generation model. This pattern was observed at the sample level as well. Panel (b) of Figure 9 presents the empirical and pseudo empirical power rates for rejecting equivalent factor means between groups. For conditions with the same sample size and general factor loadings, power increased as more indicators loaded on group factors because of the increased bias in factor mean differences.

The positive relationship between bias and the number of indicators loading on a group factor is consistent with the findings for assessing scalar invariance for the single-factor analysis models. In assessing scalar invariance, fit indices were less sensitive to structures with more indicators loading on a group factor. This is because mean differences in the ignored group factors were manifested as mean differences in the single factor rather than differences in the intercepts when all indicators had homogenously greater means in one group versus the other group. As a result, for structures 3 and 4, misfit of the scalar invariant model was smaller, but bias in estimates of factor means was greater and thus leading to greater power rates.

Standardized effect sizes for testing factor mean differences. A standardized effect size measure was calculated using the estimate of factor mean difference and the pooled factor variance estimates. Table 4 presents the standardized effect sizes for fitting the bifactor and the single-factor models at the population level. The standardized effect size measure had the same pattern with the bias in factor mean differences across

generation and analysis conditions. The only difference is that the effect sizes were smaller for conditions with stronger general factor loadings (λ_{GEN} =.7 versus λ_{GEN} =.5).

In the last column of Table 4, I calculated the differences between the effect sizes for the bifactor analysis models and the effect sizes for the single-factor analysis models. This difference reflects the additional bias in estimating the latent means introduced by fitting an analysis model that was misspecified in terms of unidimensionality. Given moderate to large effect sizes for the group factor means in the generated data, the additional bias introduced by misspecification was under .10 if only one group factor with three indicators was present. As more indicators loaded on a group factor for a generation model, bias increased substantially for a single-factor analysis model. For data generated with all indicators loading on a group factor, misspecifying unidimensionality sometimes lead to bias in estimates of factor mean difference greater than .20.

Summary

This section briefly summarizes the results of fitting bifactor and single-factor models to bifactor data in assessing MI and evaluating factor mean differences.

Not surprisingly, bifactor models that were properly specified in terms of between-group equality constraints fit the bifactor data well at the sample and population levels. However, at the sample level, the models failed to converge for a non-trivial number of replications across all generation conditions. When bifactor models were improperly specified in terms of between-group equality constraints, the models tended to have the convergence problem not only at the sample level, but also at the population level for some conditions. At the sample level, the fewest numbers of replications with properly converged models were for generation conditions with all indicators associated

with group factors and with different numbers of indicators across group factors. For the same generation conditions, the models failed to converge properly at the population level. I postulated that this convergence problem led to the large variabilities in the empirical error rates for assessing metric invariance and scalar invariance across the factor structures. Provocatively, bias in estimating mean differences on the general factor based on the bifactor analysis models were minimal across conditions.

Fitting single-factor models to bifactor data led to model misfit in examining configural invariance (i.e., with no between-group equality constraints). The degree of misfit was found to be a function of the bifactor structure underlying the data. Based on the three fit indices, the single-factor analysis models evidenced insufficient fit for data generated based on two of the four bifactor structures (structures 3 and 4). These two bifactor structures had more group factors and fewer indicators per group factor. In assessing metric invariance, the empirical Type I error rates for the single-factor analysis models had less variability, and the power rates were greater compared to the bifactor analysis models. I hypothesized that these results were due to the greater number of replications that properly converged in each analysis condition. For the single-factor analysis models, power for rejecting metric invariance increased as the number of indicators per group factor increased. In testing scalar invariance, the empirical power rates differed as a function of the number of indicators associated with a group factor in the generation factor structure. Scalar invariance tended to be rejected for single-factor analysis models when a small number of indicators loaded on a group factor with different between-group factor means. The estimation biases of the factor mean differences were substantially larger when fitting single-factor models in comparison

with bifactor models. In contrast with assessing scalar invariance, bias in estimates of mean differences and pseudo power for rejecting equivalent factor means were greater if more indicators were associated with a group factor.

Testing latent mean differences between groups requires the establishment of MI or partial MI, although the latter is not explored in this study. Failure to achieve a certain level of MI can lead to biased estimates of latent means if the incorrect invariance constraints are maintained. One objective of the study is to examine if the results of assessing MI can be informative in indicating the size of bias in the estimation of latent mean differences. Figures 10 and 11 plot the relationships between the results of assessing MI at the different hierarchical levels and the size of bias in latent mean difference estimates when fitting single-factor analysis models. Figure 10 shows the relationships at the population level, and Figure 11 illustrates the relationships at the sample level.

Figures 10a and 11a depict the relationship between model fit for assessing configural invariance and bias in estimation. Model misfit for assessing configural invariance based on the three fit indices is positively related with estimation bias across generation conditions. Generation conditions with greater bias in the estimates of latent mean differences yielded worse fit when fitting single-factor models. For conditions with bias greater than .10, RMSEA and CFI indicated inadequate model fit for the single-factor models using conventional cutoff criteria (RMSEA greater than .06 and CFI smaller than .95; Hu & Bentler, 1999), whereas SRMR suggested adequate fit at both the population (all below .04) and the sample levels (all below .05). The relationship between model misfit for assessing configural invariance and estimation bias is moderated by the

generation factor structure. Holding the general factor loadings constant, factor structures with worse model fit tended to have greater estimation bias with the exception of structure 1. The positive relationship between model misfit for assessing configural invariance and estimation bias became much weaker within factor structures.

Figures 10b and 11b show the relationship between the changes in model fit for assessing scalar invariance and bias in estimates of latent mean differences. Provocatively, model misfit for assessing scalar invariance is associated with a lack of bias in latent mean estimates. More specifically, for conditions with substantial bias (bias greater than .10), minimal model misfit is observed when imposing invariance constraints on between-group intercepts. This result is consistent across all fit indices at both population and sample levels. Similarly, power rates for detecting scalar noninvariance were very low for conditions with substantial bias. In addition, RMSEA and CFI were found to be sensitive to a lack of scalar invariance for conditions with relatively small bias. At the sample level, the increases in RMSEA when fitting a scalar invariant model to a metric invariant model were all greater than .003 for generation conditions under structure 1. The decreases in CFI were all greater than .004 for the same conditions. Similar to the results for assessing configural invariance, this relationship between the fit in assessing scalar invariance and estimation bias was moderated by the bifactor structures underlying data. With the general factor loadings and sample size held constant, factor structures with smaller changes in fit and low power rates tended to have greater bias in estimation. This relationship was found to be much weaker within each factor structure.

CHAPTER 5

DISCUSSION

Assessment of MI starts with testing whether a presumed factor structure holds for multiple populations. A unidimensional structure is tested most commonly. In practice, however, tests and scales are unlikely to yield strictly unidimensional data and are likely to be affected by one or more secondary dimensions. The effects of misspecifying the secondary dimensions on model estimation have been extensively investigated in analyses with a single population (e.g., Bonifay et al., 2015; Drasgow & Parsons, 1983; Kirisci et al., 2001; Reckase, 1979; Reise et al., 2013). In contrast, the effects of such misspecifications on MI assessments have not been studied. This study explored the effects of including or ignoring the secondary factors in analysis models on assessing MI and testing factor mean differences using a bifactor modeling framework.

The results of the study showed it is important to analyze bifactor data with an appropriate bifactor model in multiple group analysis. When the analysis model was consistent with the generation model, estimates of the latent mean differences were unbiased or slightly biased.

The drawback of applying a bifactor model is that these models are more complex and thus are less likely to converge properly. Compared to single-factor models, bifactor models have more parameters to estimate for the same set of indicators and thus require a larger sample size for proper model convergence. The current study used sample sizes of 150 and 300 for each group, and the highest convergence number was 810 out of 1000 replications for a $N = 300$ condition. To avoid non-convergence in practice, a larger sample size would be useful for analyzing bifactor models in assessing MI.

The literature suggests that proper convergence of factor models depends on the magnitudes of the communalities of indicators and the number of indicators per factor in addition to sample size (Gagne & Hancock, 2006; MacCallum, Widaman, Zhang, & Hong, 1999). The results of the bifactor analysis models at the sample level showed that conditions with higher general factor loadings had more replications with properly converged models, given the same strength of group factor loadings. These results are consistent with the literature suggesting higher convergence rates occur with stronger communalities. Counter to the literature, the number of indicators per factor did not have a uniform impact on convergence of the bifactor models. With relatively weak group factors, the numbers of properly converged replications decreased dramatically as the number of group factors increased. For the two structures with multiple group factors, the numbers of properly converged replications were lower for structure 3 than for structure 4, despite that structure 3 had more indicators per factor and had equal or more degrees of freedom than structure 4. These results suggest that scale developers should be cautious if multiple secondary dimensions are likely. In addition, the number of items falling on each of the secondary dimensions should be kept as similar as possible to avoid nonconvergence, especially when sample size is small. Applying bifactor models in practice may be aided by a priori knowledge of the true multidimensional factor structure to reach successful model convergence and satisfactory analysis results. In such cases, a Bayesian analysis approach might be considered to facilitate model estimation and to better reflect theories by incorporating substantively driven, small-variance priors for model parameters (Muthén & Asparouhov, 2012).

My study used a bifactor modeling framework to represent multidimensional factor structures. Bifactor models have been advocated in recent studies to characterize multidimensional data (e.g., Gibbons et al., 2008; Reise et al., 2007), and have been found to be helpful in the interpretation of psychological constructs (Chen, Hayes, Carver, Laurenceau, & Zhang, 2012). In the research literature on MI, however, few studies have considered a bifactor structure to represent multidimensional data, although the lack of MI is often conceptualized as the presence of multidimensionality. In my study, the lack of MI is conceptualized as between-group differences in indicator weights on the group factors and between-group differences in latent means of the group factors. The group factors represent content or context based factors that are independent of the general factor. One finding of my study is using single-factor models to analyze bifactor data is likely to confound differences in indicator parameters with difference in means of the latent distribution of interest. Depending on the generation bifactor structure, mean differences in the group factors produced either differences in indicator intercepts or differences in factor means in the single-factor models. With only a small number of indictors loading on a group factor, the mean differences in the group factors yielded differences in indicator intercepts. If detected, invariance constraints on the model can be respecified to avoid bias in latent mean comparison. As the number of indicators loading on group factors increased, the mean differences in the group factors were less likely to be detected as scalar noninvariance. The mean differences in the group factors then led to substantial bias in estimates of mean difference of the factor in the single-factor model.

The results of the study have implications for applied studies in the process of constructing measures, selecting samples, and choosing statistical models. First, if the

goal of a multiple group study is to compare population means on a single latent variable, researchers should try to minimize the number of indicators that can potentially be affected by secondary factors. If the inclusion of items with similar content or context is inevitable, one should consider items reflecting many aspects of the construct, with each aspect containing only a small number of items, such as the bifactor structure 4 in this study as opposed to structure 3.

Second, when selecting samples, one should try to minimize group differences on the possible secondary dimensions. Group comparison on the primary factor is not affected as long as the two groups have identical distributions on the secondary factors.

Lastly, alternative factor structures should be considered when applying statistical models before conducting any MI analysis. The analysis models can include a single-factor model and potentially different bifactor models with variations in the group factor structures. When comparing the fit of a single-factor model with a bifactor model, applied researchers should be cautious in that the fit function value for a bifactor model will yield a value smaller than or equal to the value for a single-factor model because the latter is nested within the former. The additional parameters of the bifactor model can capitalize on the idiosyncratic features of the sample to yield fit statistic that produce too positive view of the model, particularly if researchers conduct exploratory methods to specify the bifactor model. Thus, it is important that the selection of bifactor models is based on strong substantive theory and/or can be cross-validated to avoid overfitting. For cases where the fit of one model is not superior to the fit of the other, one might consider assessing MI using both bifactor and single-factor models and assess the results of the two models at each of the steps in MI assessment. A stronger conclusion can be reached

if the results are comparable across both bifactor and single-factor models. Moreover, if the MI results are comparable, it seems more likely that the estimated mean difference on the primary factor would be the similar across the two types of models.

There are a number of ways that future researchers could expand on my findings. As with any Monte Carlo study, it was impossible to design a study that allowed me to reach conclusions across all possible conditions. In future studies, researchers should investigate additional dimensions. For example, it would be interesting to manipulate the number of indicators per group factor, independent of the number of group factors in the generation of the data. Also, the total number of indicators for factor structures could be varied, as well as the relative magnitude of the group factor loadings to the general factor loadings. With my generation models, only the means of the latent distributions for the group factors were manipulated across groups. Different levels of variances (as well as other moments about the mean) of the latent distributions should also be considered. Finally, it would be beneficial to understand the effect of analyzing bifactor-generated data with a single-factor model when assessing MI on not only bias of differences in the general factor means, but also bias of differences in structure coefficients relating the general factor to external correlates.

REFERENCES

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, *13*, 113–127.

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of educational measurement*, *29*(1), 67-91.

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137-143.

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

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 504-516.

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

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling*, *13*(2), 287-321.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement*, *16*(2), 129-147.

Carle, A. C., Millsap, R. E., & Cole, D. A. (2007). Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent variable models. *Educational and Psychological Measurement*.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural equation modeling*, *14*(3), 464-504.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of personality and social psychology*, *95*(5), 1005.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J.-P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality, 80*, 219-251.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research*, *41*(2), 189-225.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233-255.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin*, *114*(1), 174.

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155-170.

DeMars, C. E. (2006). Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of Educational Measurement*, *43*(2), 145-168.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, *7*(2), 189-199.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Psychology Press.

Fan, X., & Sivo, S. A. (2009). Using Δgoodness-of-fit indexes in assessing mean structure invariance. *Structural Equation Modeling*, *16*(1), 54-69.

French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378-402.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., ..., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*.

Green, S. B., & Thompson, S. M. (2012). A flexible structural equation modeling approach for analyzing means. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 393-416). New York: Guilford Press.

Green, S., & Yang, Y. (2017). Empirical Underidentification with the Bifactor Model: A Case Study. *Educational and Psychological Measurement*, 0013164417719947.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*(3), 373-388.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2000). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 534-556.

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139-164.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144.

Jak, S., Oort, F. J., & Dolan, C. V. (2010). Measurement bias and multidimensionality; an illustration of bias detection in multidimensional measurement models. *AStA Advances in Statistical Analysis*, *94*(2), 129-137.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*(1), 32-60.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The role of referent indicators in tests of measurement invariance. *Structural Equation Modeling*, *16*(4), 642-657.

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

Kaplan, D., & George, R. (1995). A study of the power associated with testing factor mean differences under violations of factorial invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *2*(2), 101-118.

Kok, F. (1988). Item bias and test multidimensionality. In *Latent trait and latent class models* (pp. 263-275). Springer US.

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied psychological measurement*, *25*(2), 146-162.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First-and higher order factor models and their invariance across groups. *Psychological bulletin*, *97*(3), 562.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100-117.

McDonald, R. P. (1999). Test theory: A unified approach. *Mahwah, NJ: Lawrence Earlbaum*.

McDonald, R. P., & Mok, M. M. C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*(1), 23-40.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568.

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

Meade, A. W., & Tonidandel, S. (2010). Not seeing clearly with Cleary: What test bias analyses do and do not tell us. *Industrial and Organizational Psychology*, *3*(2), 192-205.

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

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248.

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

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological methods*, *9*(1), 93.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2015). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-24.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological methods*, *17*(3), 313.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics*, *18*(1), 41-68.

Oshima, T. C., & Miller, M. D. (1990). Multidimensionality and IRT-Based Item lnvariance Indexes: The Effect of Between-Group Variation in Trait Correlation. *Journal of Educational Measurement*, *27*(3), 273-283.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics*, *4*(3), 207-230.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*(5), 667-696.

Reise, S. P., Cook, K. F., Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In Reise, S. P., Revicki, D. A. (Eds.), *Handbook of item response theory modeling* (pp. 13-40). New York, NY: Taylor & Francis.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544-559.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures.*Quality of Life Research*, *16*(1), 19-31.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective. *Educational and Psychological Measurement*, *73*(1), 5-26.

Roussos, L. A., & Stout, W. F. (1996). Simulation Studies of the Effects of Small Sample Size and Studied Item Parameters on SIBTEST and Mantel-Haenszel Type I Error Performance. *Journal of Educational Measurement*, *33*(2), 215-230.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159-194.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229-239.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-107.

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*(1), 81.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52, 589−617.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics: Psychometrics* (pp. 683-718). Amsterdam: Elsevier.

Ten Berge, J. M., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613-625.

Thompson, M. S. & Green, S. B. (2013). Evaluating between-group differences in latent variable means. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp.163-218). lap.

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

Wang, D., Whittaker, T. A., & Beretvas, S. N. (2012). The impact of violating factor scaling method assumptions on latent mean difference testing in structured means models. *Journal of Modern Applied Statistical Methods*, *11*(1), 3.

Whittaker, T. A. (2013). The Impact of Noninvariant Intercepts in Latent Means Models. *Structural Equation Modeling*, *20*(1), 108-130.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. *The science of prevention: Methodological advances from alcohol and substance abuse research*, 281-324.

Xu, Y., & Green, S. B. (2015). The Impact of Varying the Number of Measurement Invariance Constraints on the Assessment of Between-Group Differences of Latent Means. *Structural Equation Modeling: A Multidisciplinary Journal*, 1-12.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of educational measurement*, *30*(3), 187-213.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*(3), 435-463.

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123-133.

APPENDIX A

RESULTS OF THE PRELIMINARY STUDY

Table A1

Fit indices for Fitting Single-Factor Models to Single-Group Data Generated Based on Bifactor Structure 1 with 3 Indicators Loading on One Group Factor at the Population Level.

Generation Parameter		Fit Indices		
General Factor Loading	Group Factor Loading	RMSEA	CFI	SRMR
.8	.2	.027	.997	.008
.8	.3	.060	.985	.017
.8	.4	.108	.956	.031
.8	.5	.171	.902	.065
.7	.2	.019	.998	.008
.7	.3	.042	.988	.017
.7	.4	.074	.965	.030
.7	.5	.115	.923	.050
.6	.2	.015	.997	.007
.6	.3	.033	.987	.017
.6	.4	.058	.963	.029
.6	.5	.088	.924	.047
.5	.2	.012	.997	.007
.5	.3	.028	.984	.017
.5	.4	.048	.955	.029
.5	.5	.072	.914	.045
.4	.2	.011	.995	.007
.4	.3	.024	.975	.016
.4	.4	.041	.937	.028
.4	.5	.058	.898	.041

Table A2

Fit indices for Fitting Single-Factor Models to Single-Group Data Generated Based on Bifactor Structure 2 with 6 Indicators Loading on One Group Factor at the Population Level.

Generation Parameter		Fit Indices		
General Factor Loading	Group Factor Loading	RMSEA	CFI	SRMR
.8	.2	.025	.998	.007
.8	.3	.050	.991	.016
.8	.4	.078	.981	.029
.8	.5	.103	.975	.043
.7	.2	.017	.998	.007
.7	.3	.036	.992	.016
.7	.4	.056	.984	.027
.7	.5	.074	.977	.039
.6	.2	.014	.998	.007
.6	.3	.028	.992	.015
.6	.4	.043	.984	.025
.6	.5	.057	.978	.035
.5	.2	.011	.997	.007
.5	.3	.023	.991	.014
.5	.4	.034	.984	.022
.5	.5	.044	.980	.030
.4	.2	.009	.996	.007
.4	.3	.018	.989	.013
.4	.4	.026	.984	.019
.4	.5	.033	.983	.024

Table A3

Fit indices for Fitting Single-Factor Models to Single-Group Data Generated Based on Bifactor Structure 3 with All Indicators Loading on Two Group Factors at the Population Level.

Generation Parameter		Fit Indices		
General Factor Loading	Group Factor Loading	RMSEA	CFI	SRMR
.8	.2	.051	.990	.015
.8	.3	.108	.959	.034
.8	.4	.184	.905	.063
.8	.5	.284	.838	.103
.7	.2	.036	.991	.015
.7	.3	.078	.965	.034
.7	.4	.131	.917	.060
.7	.5	.197	.860	.096
.6	.2	.029	.991	.015
.6	.3	.062	.963	.033
.6	.4	.105	.915	.059
.6	.5	.156	.860	.092
.5	.2	.025	.988	.015
.5	.3	.053	.955	.033
.5	.4	.089	.903	.058
.5	.5	.131	.849	.089
.4	.2	.022	.982	.015
.4	.3	.047	.937	.032
.4	.4	.078	.879	.056
.4	.5	.114	.829	.085

Table A4

Fit indices for Fitting Single-Factor Models to Single-Group Data Generated Based on Bifactor Structure 4 with All Indicators Loading on Three Group Factors at the Population Level.

Generation Parameter		Fit Indices		
General Factor Loading	Group Factor Loading	RMSEA	CFI	SRMR
.8	.2	.054	.988	.015
.8	.3	.122	.945	.035
.8	.4	.220	.847	.062
.8	.5	.377	.670	.097
.7	.2	.038	.990	.015
.7	.3	.086	.954	.035
.7	.4	.153	.874	.062
.7	.5	.246	.744	.097
.6	.2	.031	.989	.015
.6	.3	.069	.951	.035
.6	.4	.122	.869	.062
.6	.5	.192	.746	.097
.5	.2	.026	.986	.015
.5	.3	.059	.937	.035
.5	.4	.104	.841	.062
.5	.5	.163	.709	.097
.4	.2	.023	.977	.015
.4	.3	.052	.906	.035
.4	.4	.093	.782	.062
.4	.5	.145	.636	.097

APPENDIX B

ADDITIONAL MODEL CONVERGENCE RESULTS

Table B1

Numbers of Replications of 1000 That Reached Proper Solutions When Fitting Bifactor Models to Data with General Factor Loadings of .7 and Group Factor Mean Differences of .4 at the Sample Level

Generation Model	λ_{GRP}	*N*	Bifactor Analysis Models			
			Configural Invariance	Metric Invariance	Scalar Invariance	All three analysis models[a]
S1: One group factor (with 3 indicators)	Invariant	150	467	700	841	394
		300	623	829	948	597
	Non-invariant	150	515	810	901	477
		300	718	936	992	699
S2: One group factor (with 6 indicators)	Invariant	150	689	883	958	645
		300	857	924	999	810
	Non-invariant	150	581	730	811	468
		300	805	876	924	718
S3: Two group factors	Invariant	150	43	160	196	10
		300	71	247	250	17
	Non-invariant	150	59	127	187	8
		300	124	120	186	18
S4: Three group factors	Invariant	150	98	417	508	70
		300	267	653	711	235
	Non-invariant	150	142	537	599	114
		300	416	849	897	386

[a]The number of replications that reached proper solutions across all three analysis models: configurally invariant model, metric invariant model, and scalar invariant model

Table B2

Numbers of Replications That Reached Proper Solutions When Fitting Bifactor Models to Data with General Factor Loadings of .5 and Group Factor Mean Differences of .4 at the Sample Level

Generation Model	λ_{GRP}	*N*	Bifactor Analysis Models			
			Configural Invariance	Metric Invariance	Scalar Invariance	All three analysis models[a]
S1: One group factor (with 3 indicators)	Invariant	150	330	576	746	267
		300	537	774	945	466
	Non-invariant	150	352	581	713	288
		300	545	831	929	508
S2: One group factor (with 6 indicators)	Invariant	150	392	641	750	302
		300	586	805	934	504
	Non-invariant	150	350	573	669	239
		300	545	716	829	431
S3: Two group factors	Invariant	150	45	152	215	6
		300	73	212	244	17
	Non-invariant	150	40	119	174	6
		300	111	151	257	14
S4: Three group factors	Invariant	150	40	199	321	19
		300	144	522	607	99
	Non-invariant	150	36	199	286	14
		300	157	544	609	118

[a]The number of replications that reached proper solutions across all three analysis models: configurally invariant model, metric invariant model, and scalar invariant model

APPENDIX C

RESULTS FOR NON-UNIFORM NONINVARIANCE CONDITIONS

Table C1

Changes in Fit Indices for Assessing Metric Invariance When Fitting Single-Factor Analysis Models to Bifactor Data Generated with Uniform and Non-Uniform Noninvariant Group Factor Loadings at the Population Level (λ_{GEN}=.7 and $\Delta\kappa_{GRP}$ = .4)

Generation Model	Noninvaraince Pattern	Single-Factor Analysis Models (Δdf = 8)		
		ΔRMSEA	ΔCFI	ΔSRMR
S1: One group factor with 3 indicators	Uniform	-.0026	-.0006	.0017
	Non-uniform	-.0030	-.0001	.0002
S2: One group factor with 6 indicators	Uniform	.0008	-.0020	.0095
	Non-uniform	.0008	-.0020	.0095
S3: Two group factors with 9 indicators	Uniform	-.0046	-.0017	.0032
	Non-uniform	-.0046	-.0013	.0041
S4: Three group factors with 9 indicators	Uniform	-.0070	< .0001	.0001
	Non-uniform	-.0063	.0001	.0002

Note. ΔGoodness-of-fit indices with underlines indicate that a metric invariant model fits better than a configurally invariant model to data. Shaded cells indicate conditions with negative variance estimates of group factors for the misspecified bifactor metric invariant models.

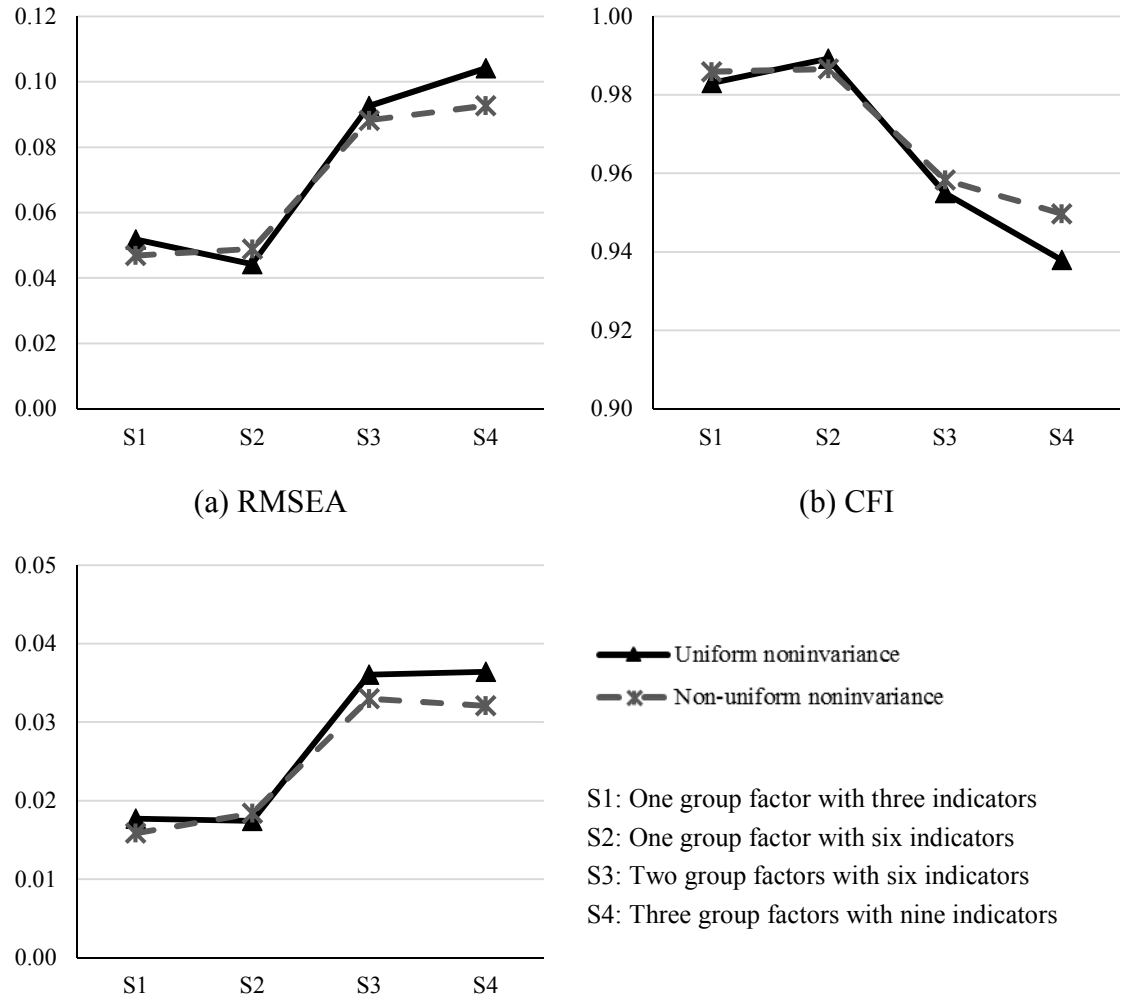


Figure C1. Fit of single-factor configurally invariant models at the population level for conditions with uniform and non-uniform noninvariant group factor loadings (λ_{GEN}=.7 and $\Delta\kappa_{GRP}$ = .4).

Table 1

Population Factor Loadings for Generation Conditions with Average General Factor Loadings of .7

	Invariant Group Factor Loading Group 1 and Group 2
Structure 1	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ .2 & .3 & .4 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
Structure 2	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ 0 & 0 & 0 & .2 & .3 & .4 & .2 & .3 & .4 \end{bmatrix}$
Structure 3	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ .2 & .3 & .4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .2 & .3 & .4 & .2 & .3 & .4 \end{bmatrix}$
Structure 4	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ .2 & .3 & .4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .2 & .3 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .2 & .3 & .4 \end{bmatrix}$

	Noninvariant Group Factor Loading	
	Group 1	Group 2
Structure 1	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.2} & .3 & \mathbf{.2} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.4} & .3 & \mathbf{.4} & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
Structure 2	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ 0 & 0 & 0 & \mathbf{.2} & .3 & \mathbf{.2} & \mathbf{.2} & .3 & \mathbf{.2} \end{bmatrix}$	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ 0 & 0 & 0 & \mathbf{.4} & .3 & \mathbf{.4} & \mathbf{.4} & .3 & \mathbf{.4} \end{bmatrix}$
Structure 3	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.2} & .3 & \mathbf{.2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{.2} & .3 & \mathbf{.2} & \mathbf{.2} & .3 & \mathbf{.2} \end{bmatrix}$	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.4} & .3 & \mathbf{.4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{.4} & .3 & \mathbf{.4} & \mathbf{.4} & .3 & \mathbf{.4} \end{bmatrix}$

Structure 4	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.2} & .3 & \mathbf{.2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{.2} & .3 & \mathbf{.2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{.2} & .3 & \mathbf{.2} \end{bmatrix}$	$\lambda' = \begin{bmatrix} .6 & .7 & .8 & .6 & .7 & .8 & .6 & .7 & .8 \\ \mathbf{.4} & .3 & \mathbf{.4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{.4} & .3 & \mathbf{.4} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{.4} & .3 & \mathbf{.4} \end{bmatrix}$

Note. Bolded loadings are group factor loadings generated to be noninvariant between groups. S1: One group factor with 3 indicators; S2: One group factor with 6 indicators; S3: Two group factors with 9 indicators; S4: Three group factors with 9 indicators.

Table 2

Changes in Fit Indices for Assessing Metric Invariance When Fitting Bifactor Analysis Models and Single-Factor Analysis Models to Bifactor Data Generated with Noninvariant Group Factor Loadings at the Population Level ($\Delta\kappa_{GRP} = .4$)

Generation Model	λ_{GEN}	Bifactor Analysis Models				Single-Factor Analysis Models (Δdf = 8)		
		Δdf	ΔRMSEA	ΔCFI	ΔSRMR	ΔRMSEA	ΔCFI	ΔSRMR
S1: One group factor with 3 indicators	.5	10	.0048	-.0005	.0028	-.0014	-.0012	.0016
	.7	10	.0070	-.0003	.0028	-.0026	-.0006	.0017
S2: One group factor with 6 indicators	.5	13	.0107	-.0018	.0078	.0025	-.0044	.0084
	.7	13	.0151	-.0013	.0085	.0008	-.0020	.0095
S3: Two group factors with 9 indicators	.5	15	.0106	-.0015	.0062	-.0025	-.0030	.0028
	.7	15	.0144	-.0010	.0059	-.0046	-.0017	.0032
S4: Three group factors with 9 indicators	.5	14	.0087	-.0012	.0047	-.0045	-.0002	.0002
	.7	14	.0133	-.0009	.0050	-.0070	< .0001	.0001

Note. Results for generation conditions with $\Delta\kappa_{GRP} = 0$ or .2 were only different from those with $\Delta\kappa_{GRP} = .4$ in the fourth decimal place; thus, only the results for $\Delta\kappa_{GRP} = .4$ were shown. ΔGoodness-of-fit indices with underlines indicate that a metric invariant model fits better than a configurally invariant model to data. Shaded cells indicate conditions with negative variance estimates of group factors for the misspecified bifactor metric invariant models.

Table 3

Changes in Fit Indices for Assessing Scalar Invariance When Fitting Single-Factor Analysis Models to Bifactor Data with Invariant Group Factor Loadings at the Population Level (Δdf = 8)

Generation Model	$\Delta\kappa_{GRP}$	λ_{GEN}	ΔRMSEA	ΔCFI	ΔSRMR
S1: One group factor with 3 indicators	.2	.5	-.0005	-.0017	.0010
		.7	-.0012	-.0012	.0011
	.4	.5	.0035	-.0077	.0039
		.7	.0031	-.0045	.0040
S2: One group factor with 6 indicators	.2	.5	-.0003	-.0011	.0010
		.7	-.0012	-.0008	.0011
	.4	.5	.0027	-.0041	.0035
		.7	.0023	-.0030	.0041
S3: Two group factors with 9 indicators	.2	.5	-.0032	.0001	< .0001
		.7	-.0052	> -.0001 & < 0	.0001
	.4	.5	-.0030	-.0003	.0001
		.7	-.0049	-.0003	.0002
S4: Three group factors with 9 indicators	.2	.5	-0.0035	> -.0001 & < 0	<.0001
		.7	-0.0057	.0003	-.0001
	.4	.5	-.0032	-.0006	.0002
		.7	-.0051	-.0005	.0002

Note. ΔGoodness-of-fit indices with underlines indicate that a scalar invariant model fits better than a metric invariant model to data.

Table 4

Bias in Estimates of Between-Group Differences in the Means and the Standardized Effect Sizes for the General/Single Factor When Fitting Bifactor and Single-Factor Scalar Invariant Models to Bifactor Data with Invariant Group Factor Loadings at the Population Level

Generation Model	$\Delta\kappa_{GRP}$	λ_{GEN}	Bifactor Scalar Invariant Models		Single-Factor Scalar Invariant Models		ΔStd Effect Size[a]
			Bias in Est	Std Effect Size	Bias in Est	Std Effect Size	
S1: One group factor with 3 indicators	.2	.5	0	0	.0230	.0471	.0471
		.7	0	0	.0240	.0347	.0347
	.4	.5	0	0	.0450	.0924	.0924
		.7	0	0	.0480	.0695	.0695
S2: One group factor with 6 indicators	.2	.5	0	0	.0390	.0879	.0879
		.7	0	0	.0450	.0690	.0690
	.4	.5	0	0	.0770	.1748	.1748
		.7	0	0	.0910	.1402	.1402
S3: Two group factors with 9 indicators	.2	.5	0	0	.0540	.1131	.1131
		.7	.0540	.0833	.0580	.0859	.0026
	.4	.5	0	0	.1070	.2236	.2236
		.7	.1080	.1668	.1160	.1718	.0050
S4: Three group factors with 9 indicators	.2	.5	0	0	.0630	.1195	.1195
		.7	0	0	.0640	.0888	.0888
	.4	.5	.0010	.0025	.1250	.2371	.2346
		.7	0	0	.1270	.1761	.1761

[a]The Δstd-effect-size measure is calculated by subtracting two standardized effect sizes: effect size of the general factor mean differences when fitting bifactor scalar invariant models and effect size of the single factor mean differences when fitting single-factor scalar invariant models.

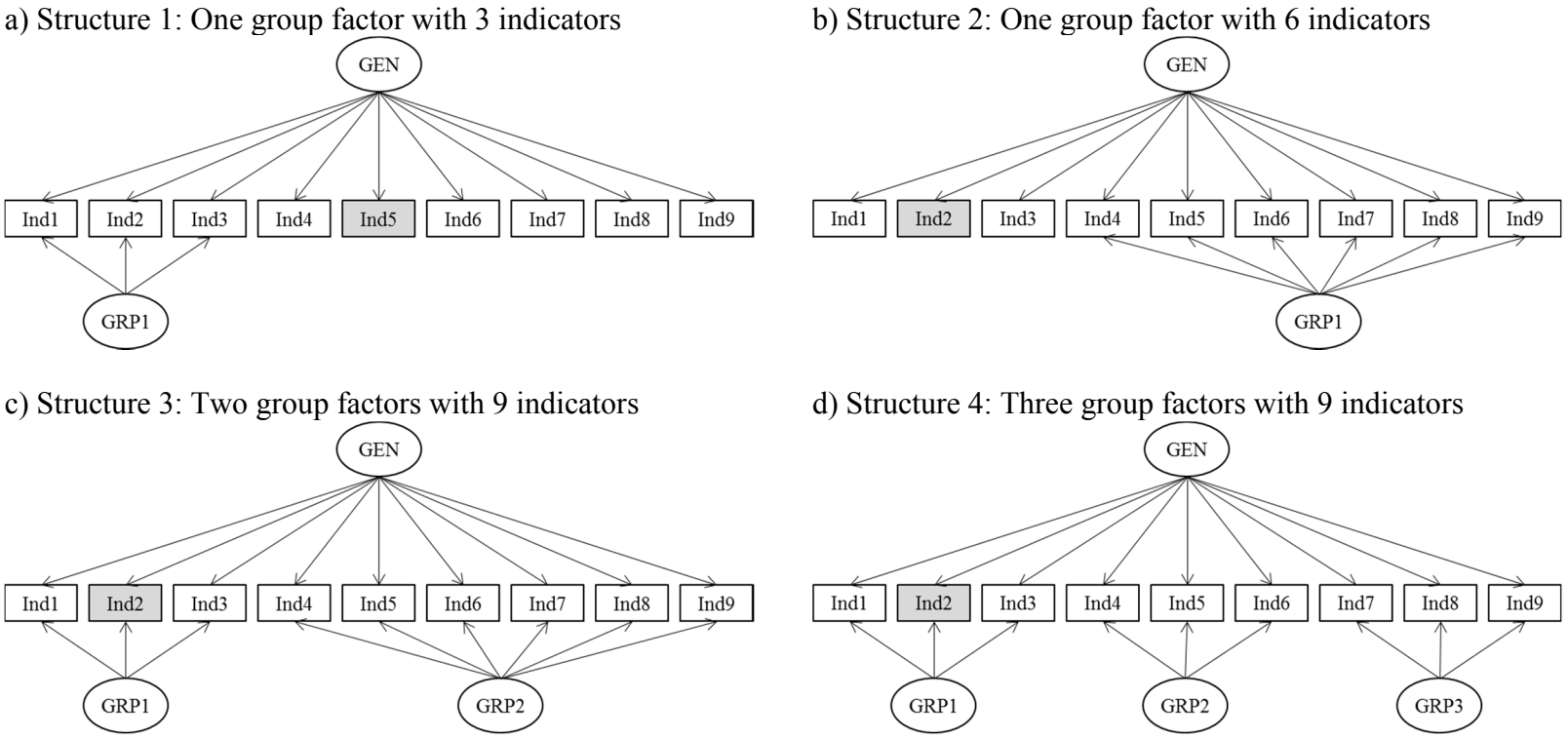


Figure 1. Path diagrams for the four bifactor structures. The shaded indicator in each diagram indicates that this indicator is used as a referent indicator when fitting configurally invariant and metric invariant single-factor models to data.

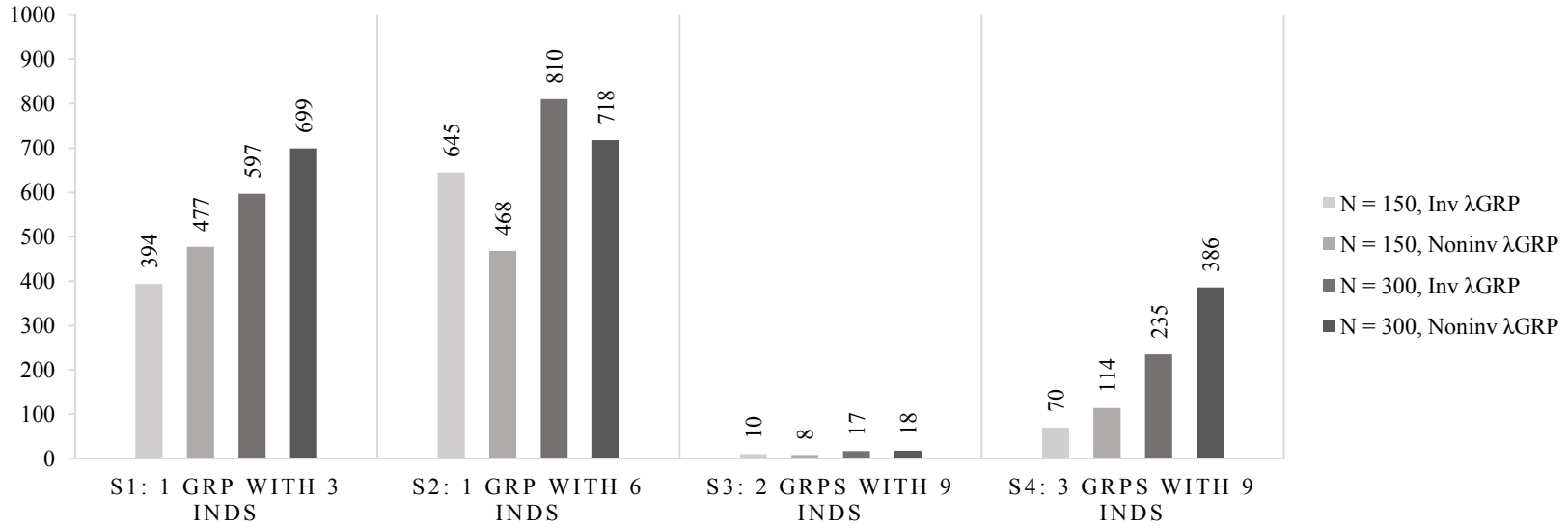


Figure 2a. Numbers of replications that reached proper solution for all three analysis models (i.e., configurally invariant model, metric invariant model, and scalar invariant model) when fitting bifactor analysis models to bifactor data at the sample level with general factor loadings of .7.

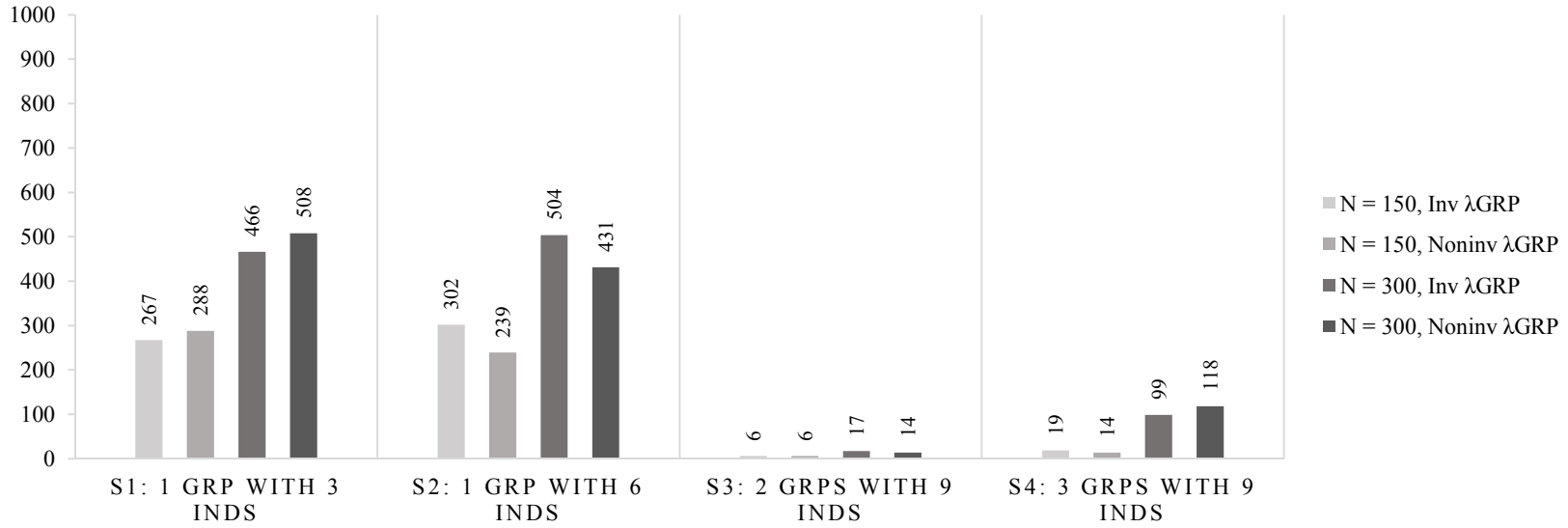


Figure 2b. Numbers of replications that reached proper solution across all three analysis models (i.e., configurally invariant model, metric invariant model, and scalar invariant model) when fitting bifactor analysis models to bifactor data at the sample level with general factor loadings of .5.

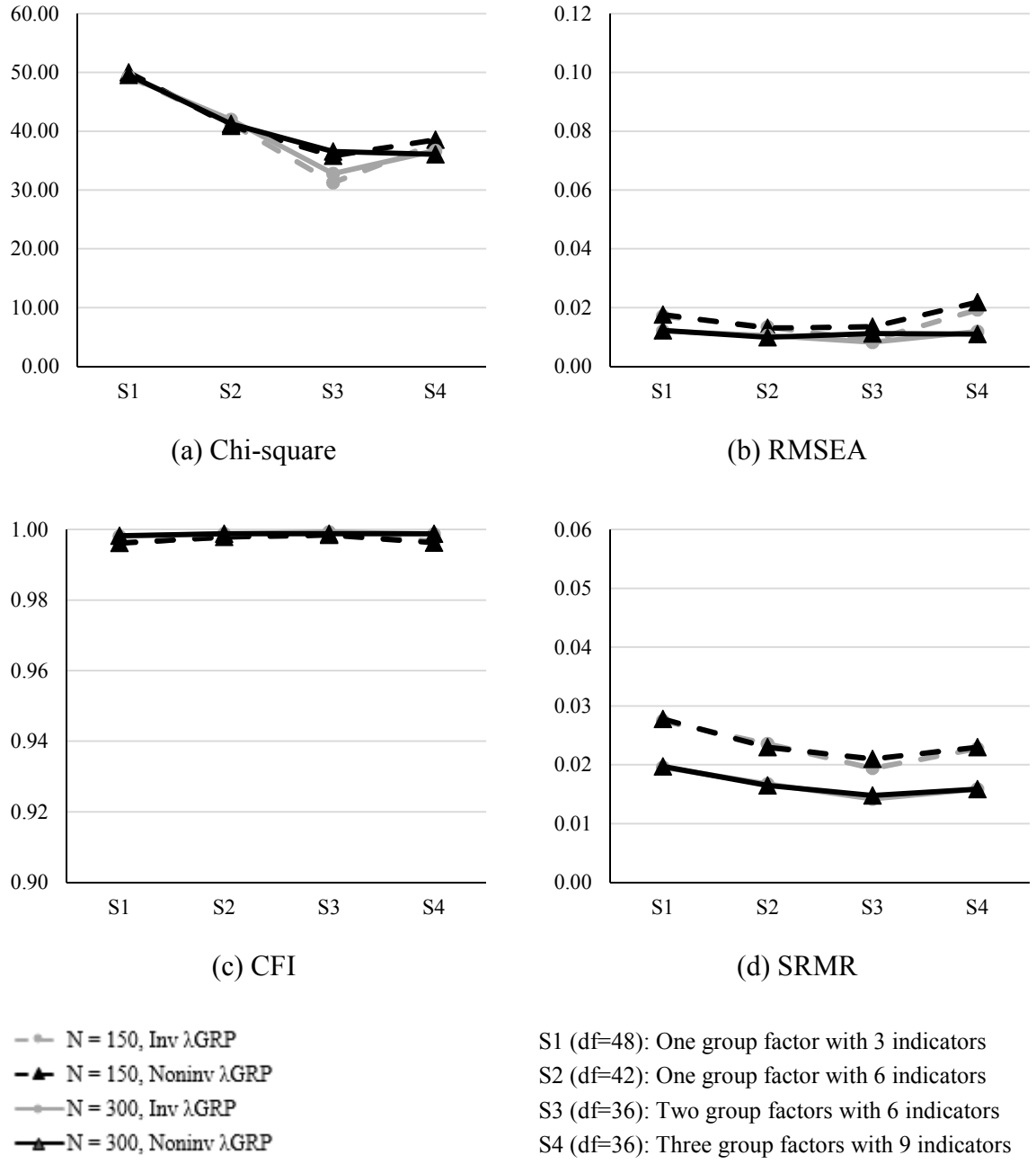


Figure 3a. Fit of bifactor factor configurally invariant models at the sample level for conditions with general factor loadings of .7 ($\Delta\kappa_{GRP} = .4$).

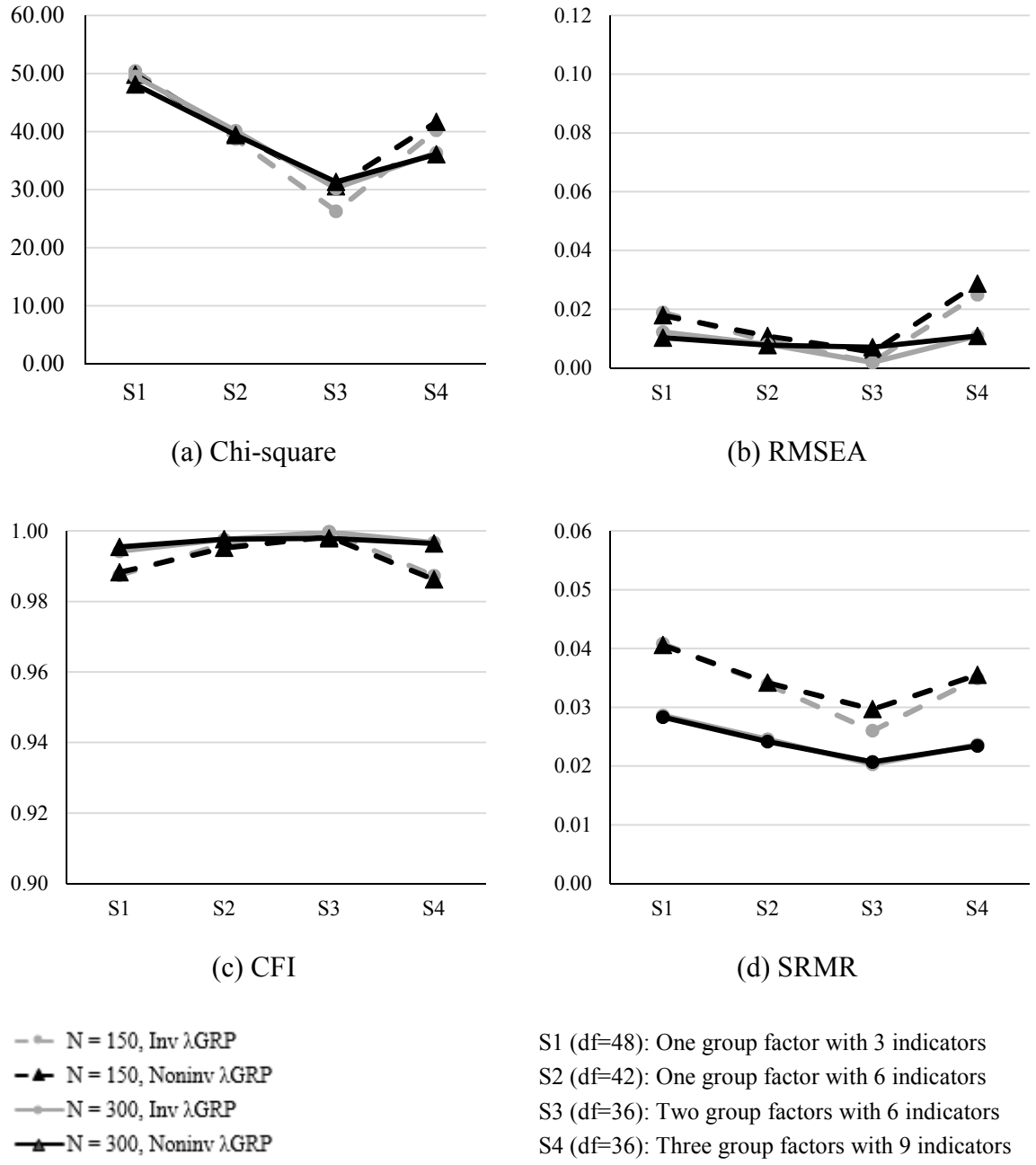


Figure 3b. Fit of bifactor factor configurally invariant models at the sample level for conditions with general factor loadings of .5 ($\Delta\kappa_{GRP} = .4$).

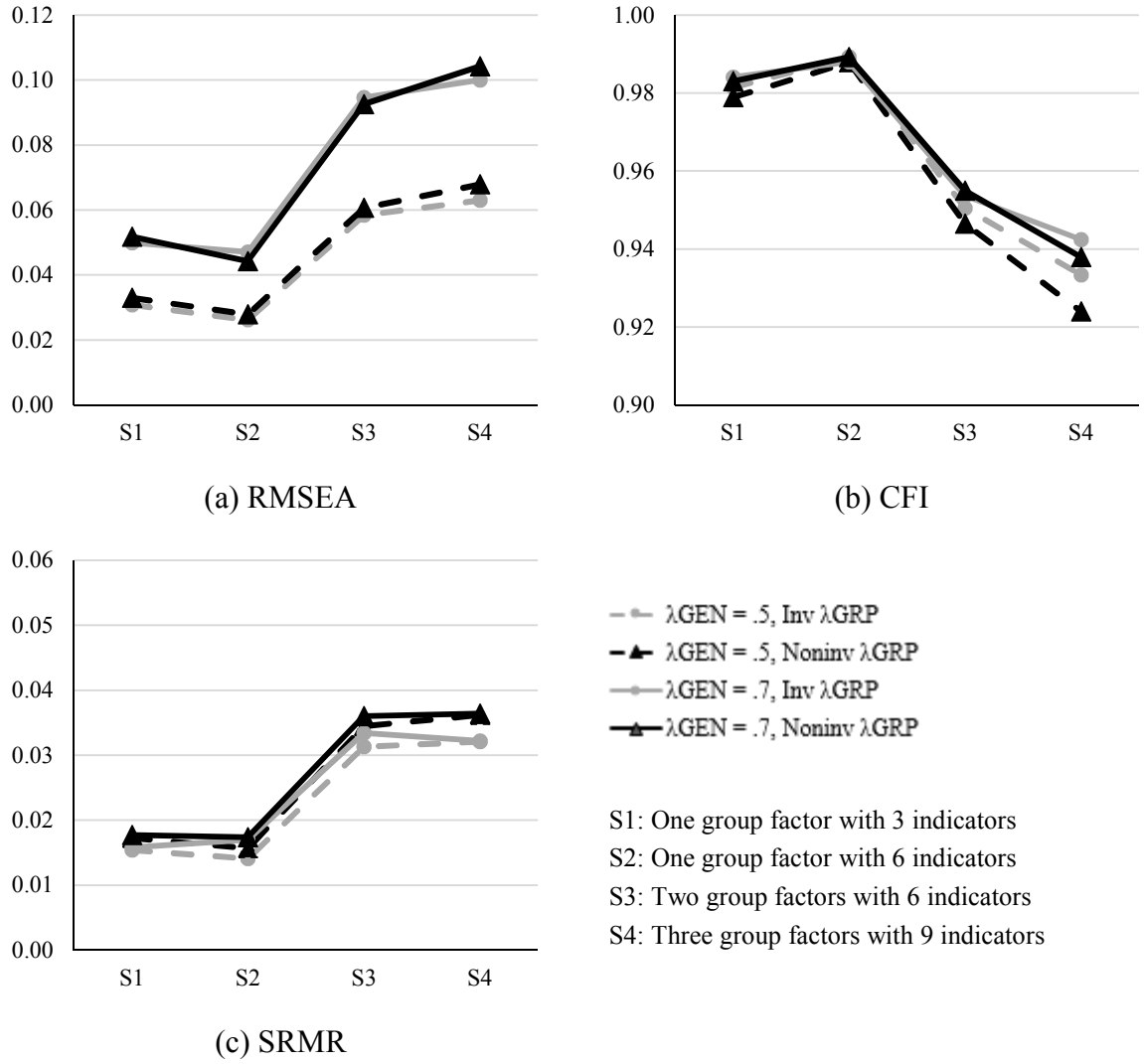


Figure 4. Fit of single-factor configurally invariant models at the population level ($\Delta\kappa_{GRP}$ = .4).

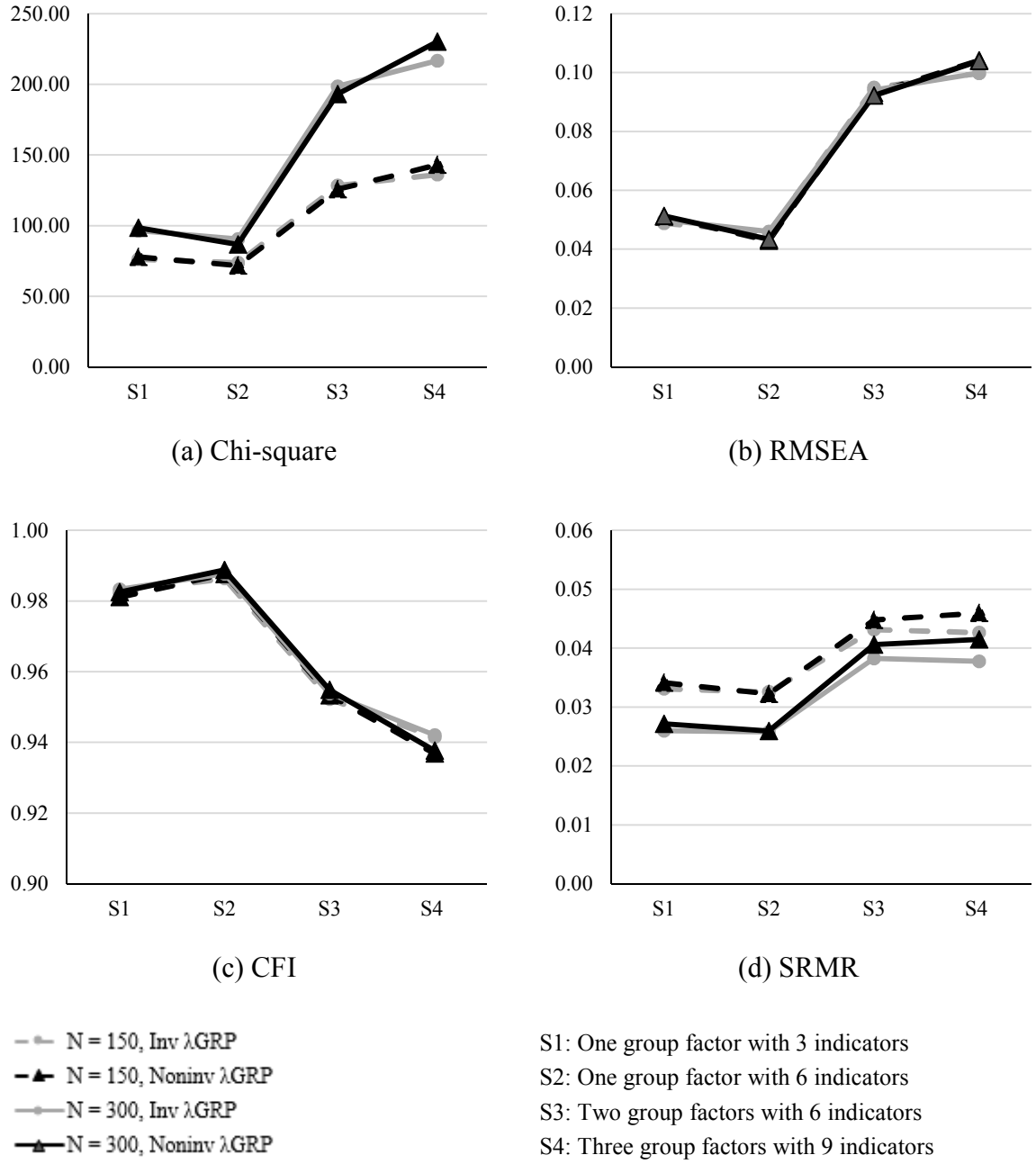


Figure 5a. Fit for single-factor configurally invariant models at the sample level for conditions with general factor loadings of .7 ($\Delta\kappa_{GRP} = .4$). Degrees of freedom of single-factor configurally invariant models fitting to all generation conditions are 54.

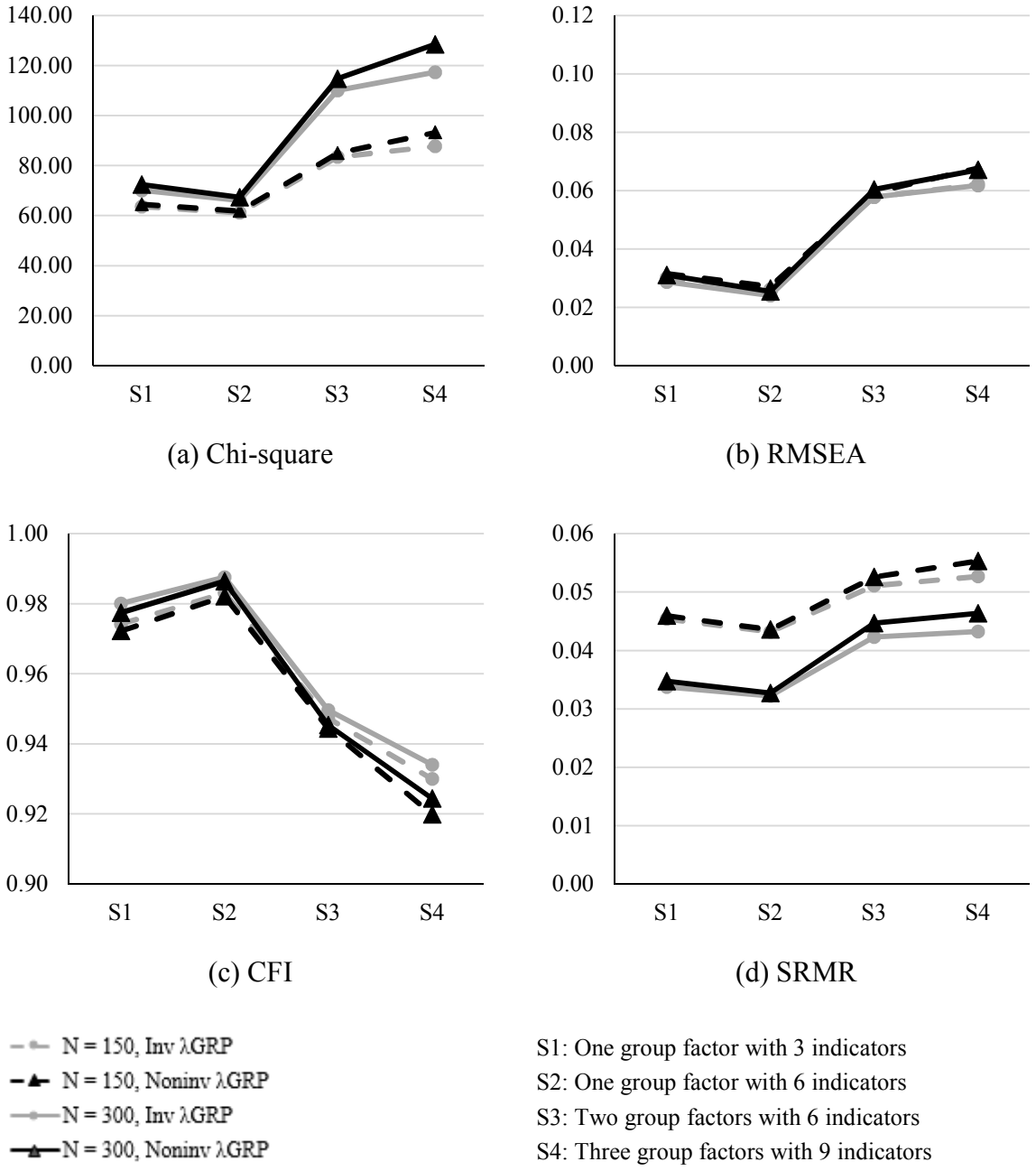
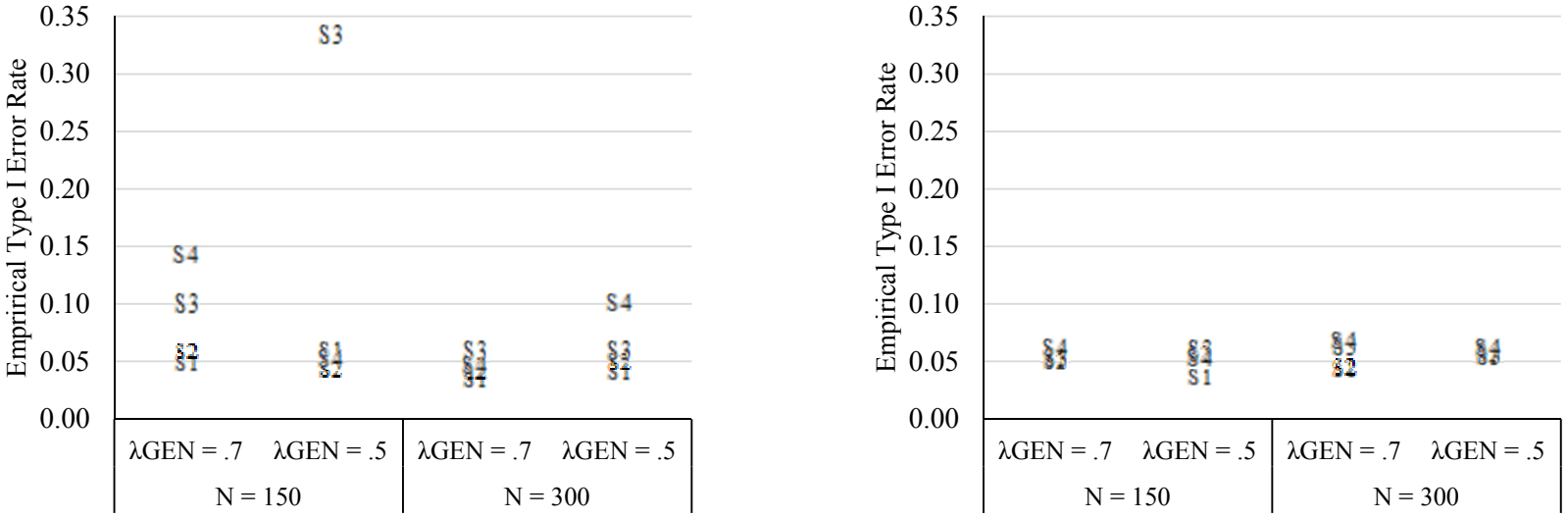


Figure 5b. Fit for single-factor configurally invariant models at the sample level for conditions with general factor loadings of .5 ($\Delta\kappa_{GRP} = .4$). Degrees of freedom of single-factor configurally invariant models fitting to all generation conditions are 54.



(a) Bifactor analysis models

(b) Single-factor analysis models

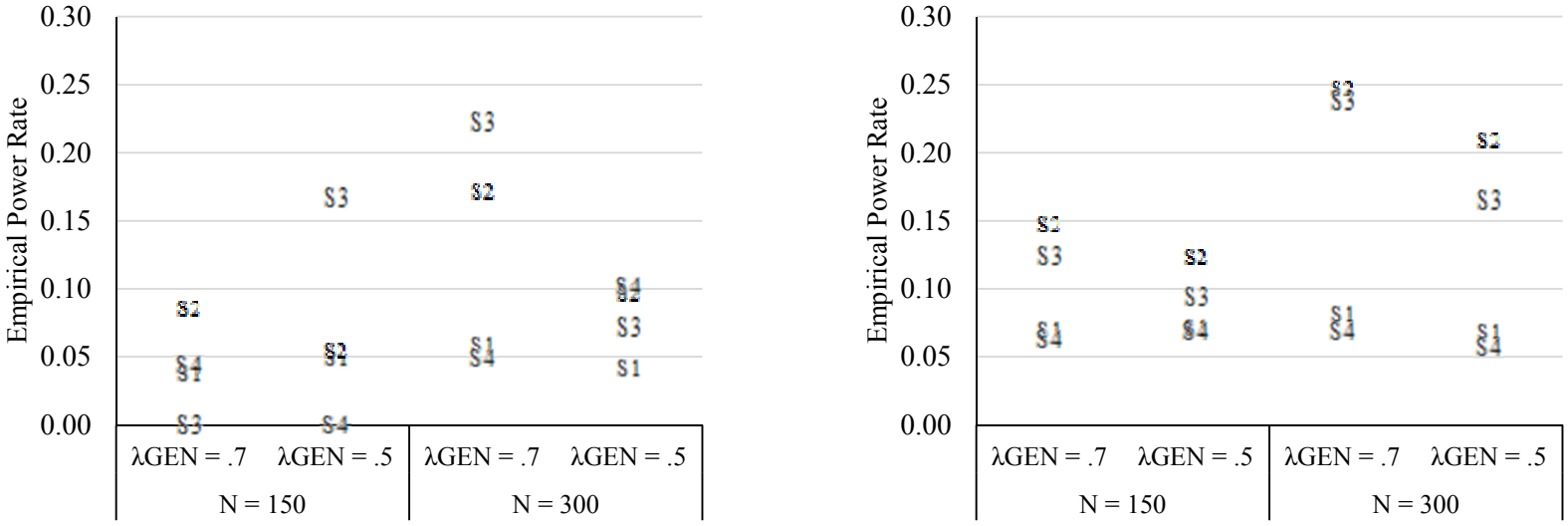
S1: One group factor with 3 indicators

S2: One group factor with 3 indicators

S3: Two group factors with 6 indicators

S4: Three group factors with 9 indicators

Figure 6. Empirical rates of rejecting a metric invariant model when fitting bifactor analysis model and single-factor analysis models to bifactor data generated with invariant group factor loadings ($\Delta\kappa_{GRP} = .4$).



(a) Bifactor analysis models

(b) Single-factor analysis models

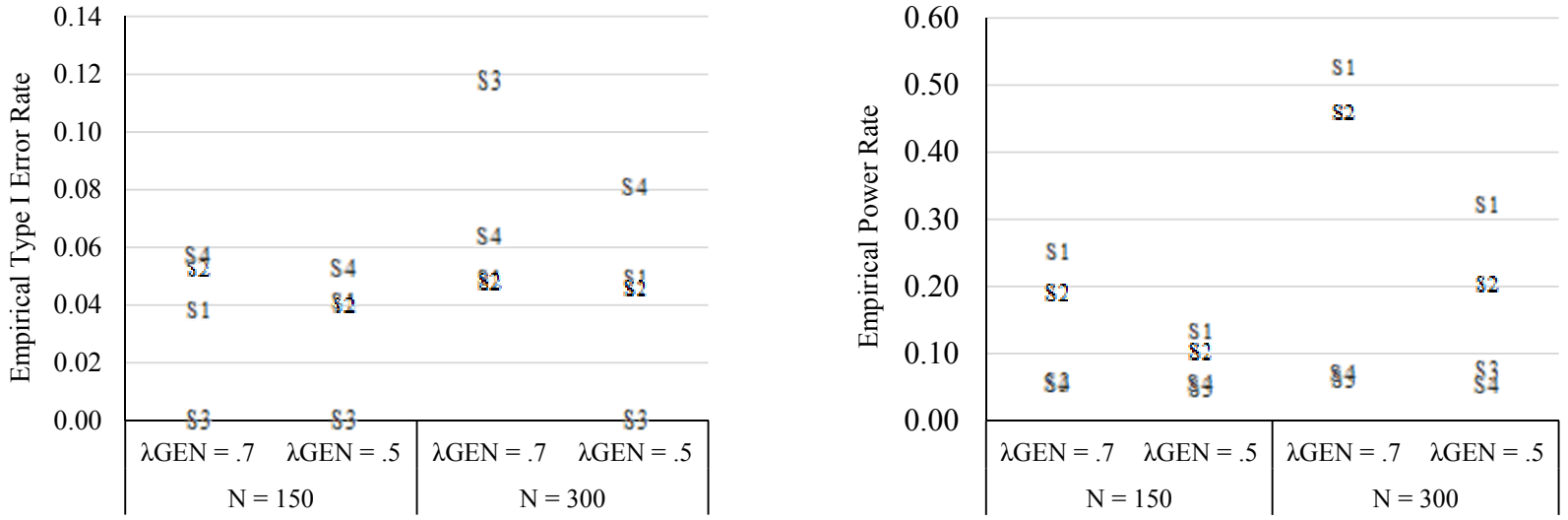
S1: One group factor with 3 indicators

S2: One group factor with 3 indicators

S3: Two group factors with 6 indicators

S4: Three group factors with 9 indicators

Figure 7. Empirical rates of rejecting metric invariance when fitting bifactor analysis model and single-factor analysis models to bifactor data generated with noninvariant group factor loadings ($\Delta\kappa_{GRP} = .4$).



(a) Bifactor analysis models

(b) Single-factor analysis models

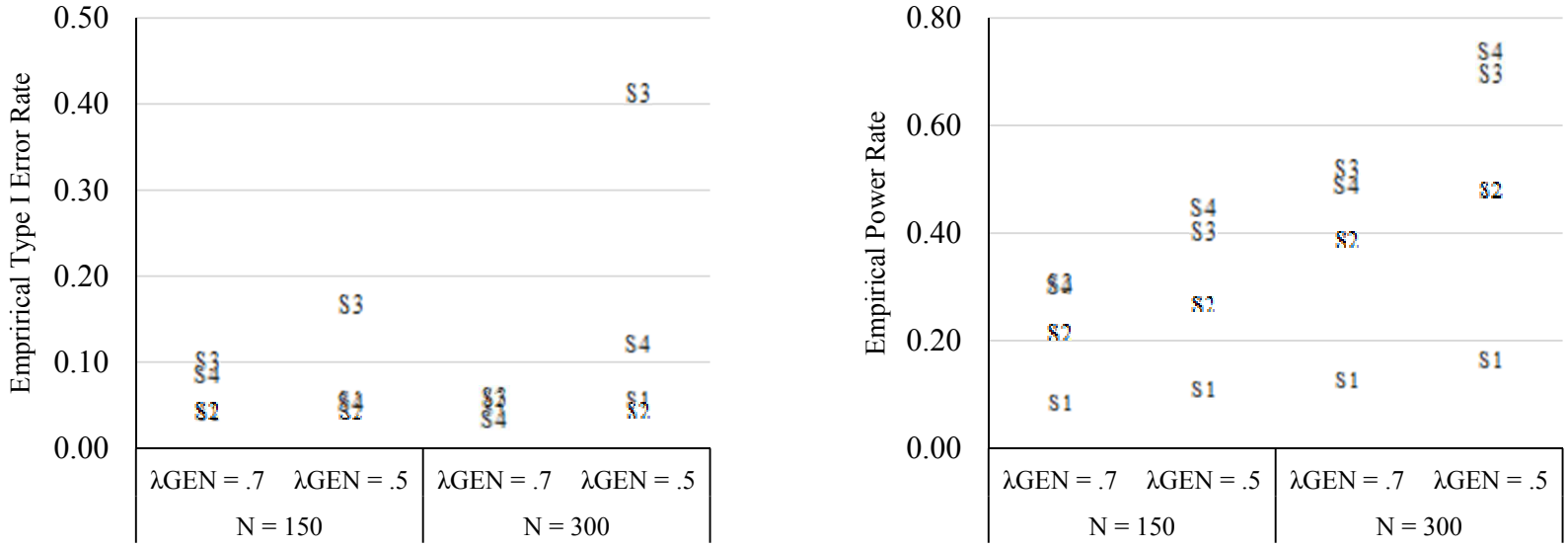
S1: One group factor with 3 indicators

S2: One group factor with 3 indicators

S3: Two group factors with 6 indicators

S4: Three group factors with 9 indicators

Figure 8. Empirical rates of rejecting scalar invariance when fitting bifactor analysis models and single-factor analysis models to bifactor data generated with invariant group factor loadings ($\Delta\kappa_{GRP} = .4$).



(a) Bifactor analysis models

(b) Single-factor analysis models

S1: One group factor with 3 indicators

S2: One group factor with 3 indicators

S3: Two group factors with 6 indicators

S4: Three group factors with 9 indicators

Figure 9. Empirical rates of rejecting equivalent between-group general/single factor means when fitting bifactor analysis models and single-factor analysis models to bifactor data generated with invariant group factor loadings ($\Delta\kappa_{GRP} = .4$).

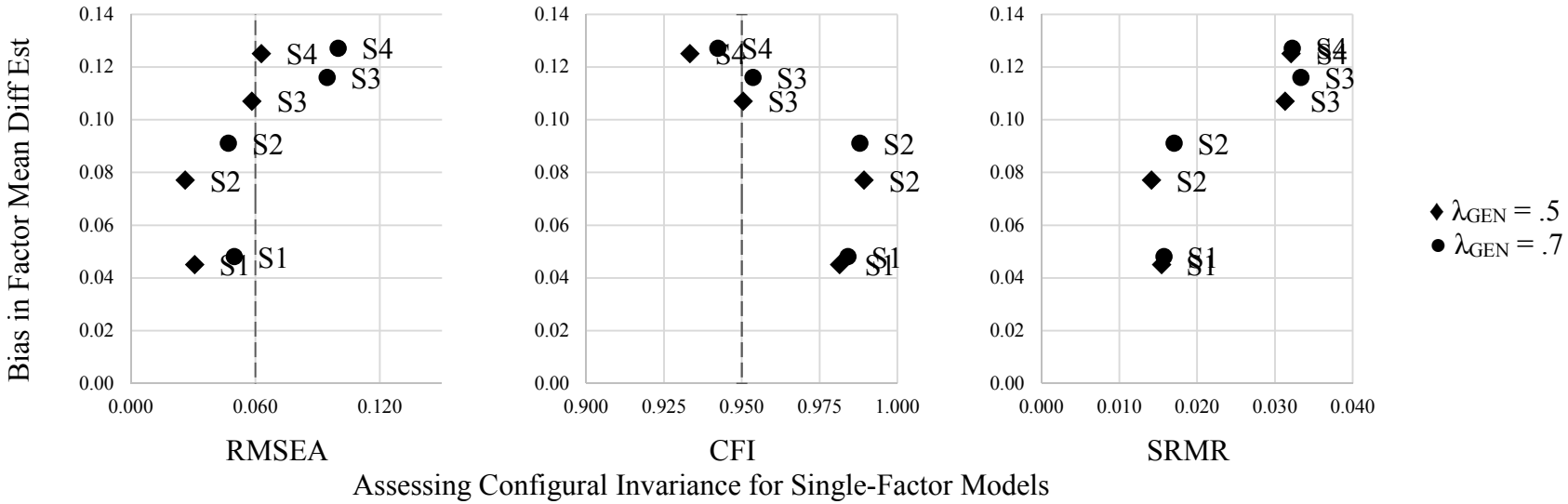


Figure 10a. Relationships between fit indices for assessing configural invariance for single-factor models and bias in estiamtes of factor mean differences at the population level ($\Delta\kappa_{GRP} = .4$).

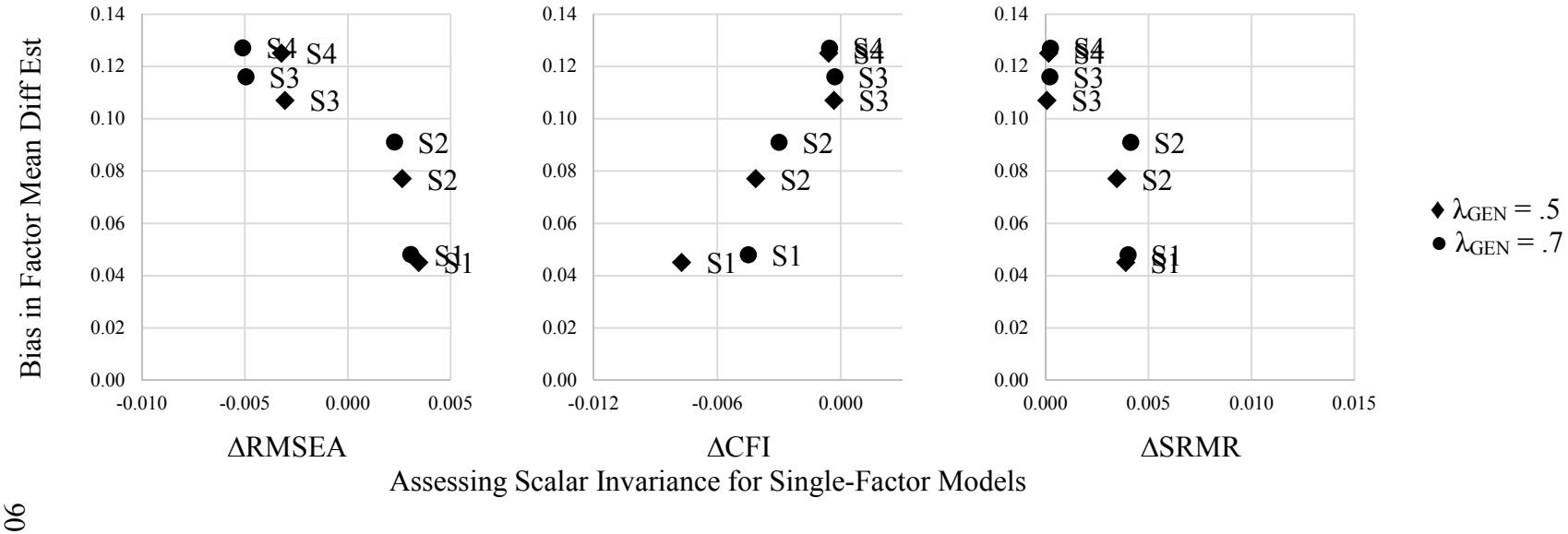


Figure 10b. Relationships between the changes in fit indices for assessing scalar invariance for single-factor models and bias in estiamtes of factor mean differences at the population level ($\Delta\kappa_{GRP} = .4$).

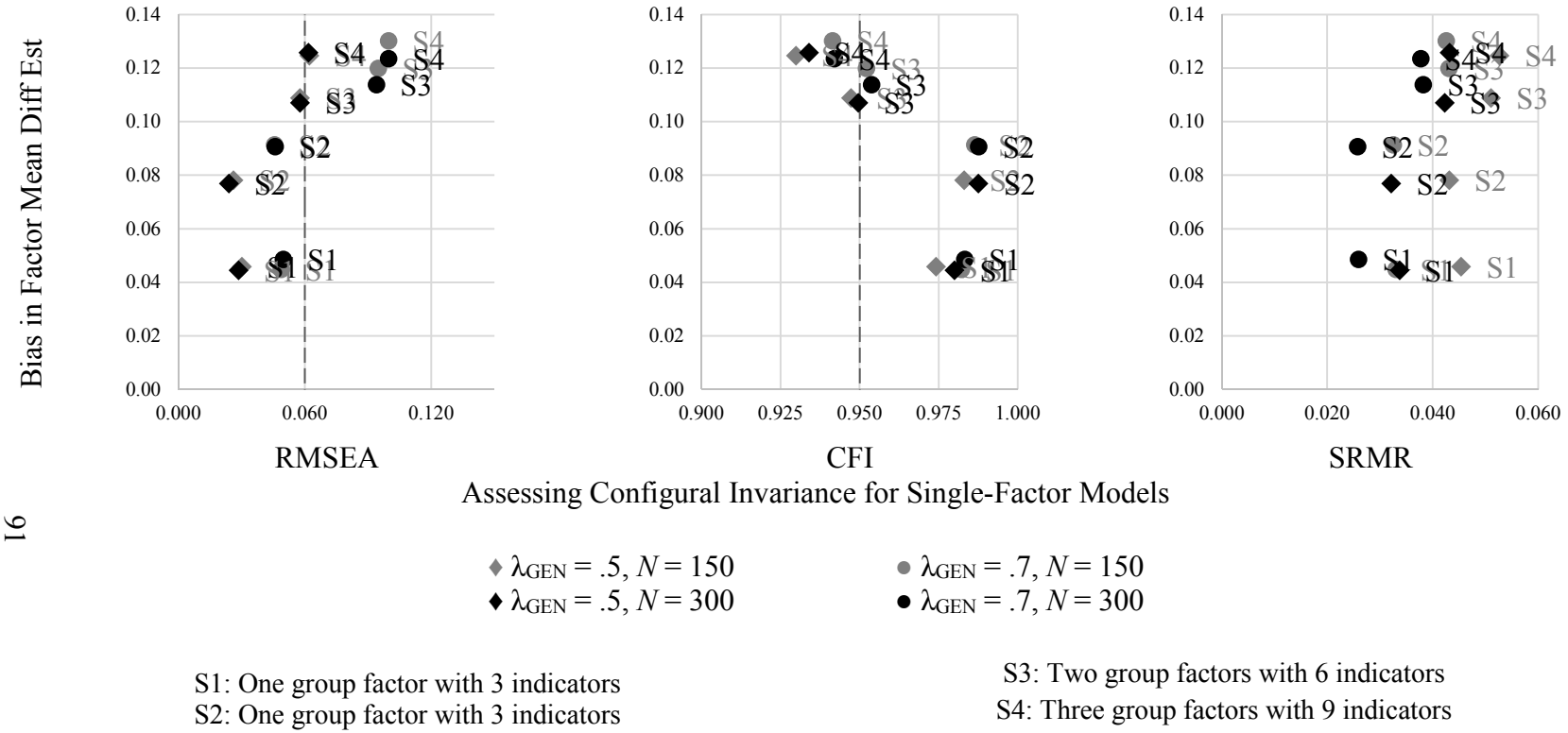


Figure 11a. Relationships between fit indices for assessing configural invariance for single-factor models and bias in estiamtes of factor mean differences at the sample level ($\Delta\kappa_{GRP} = .4$).

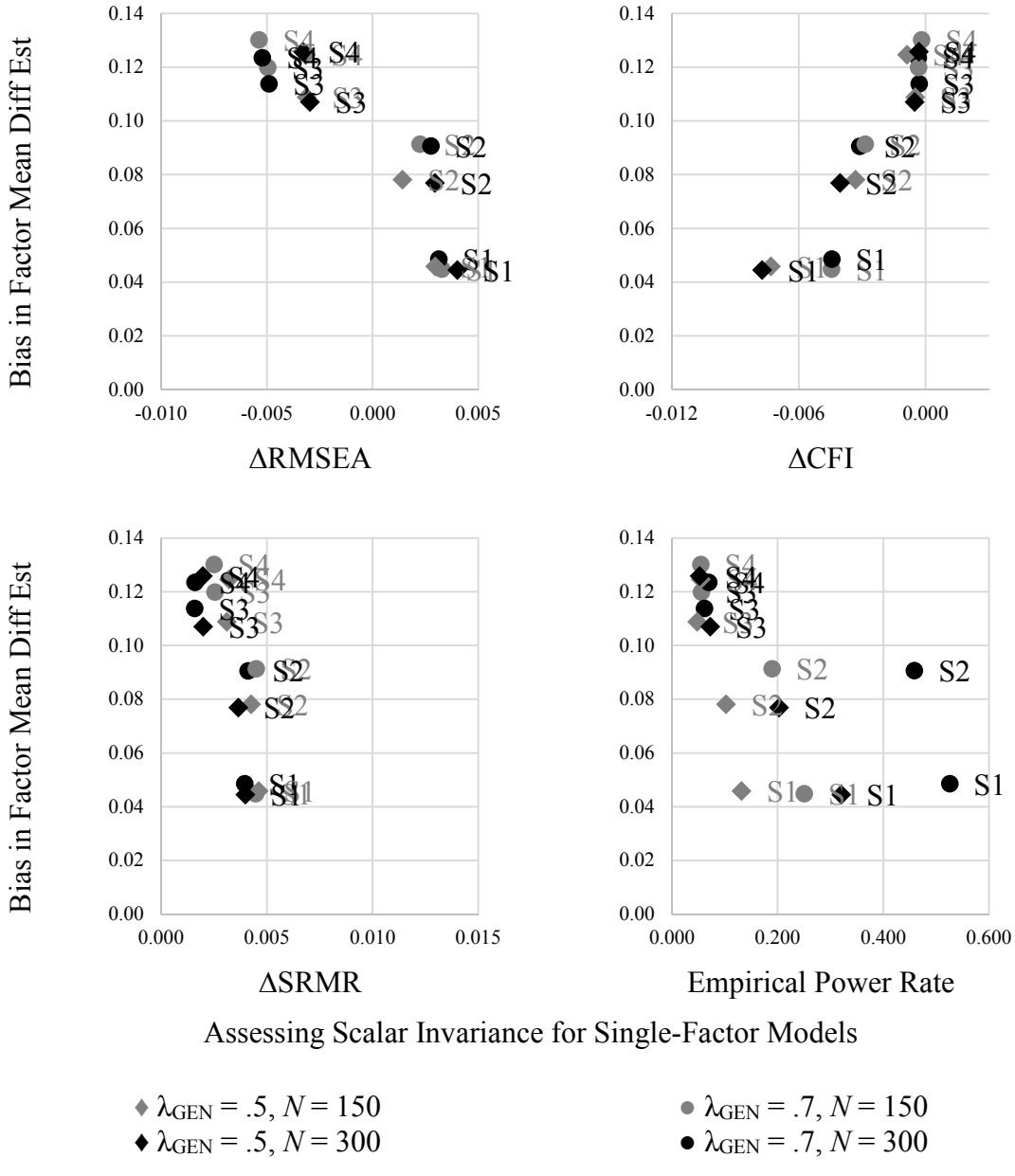


Figure 11b. Relationships between changes in fit indices for assessing scalar invariance for single-factor models and bias in estiamtes of factor mean differences at the sample level ($\Delta\kappa_{GRP} = .4$).