Robustness of the General Factor Mean Difference Estimation in Bifactor Ordinal Data

by

Yixing Liu

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

Approved April 2019 by the
Graduate Supervisory Committee

Marilyn Thompson, Chair
Roy Levy
Holly O'Rourke

ARIZONA STATE UNIVERSITY

May 2019

ABSTRACT

A simulation study was conducted to explore the robustness of general factor mean difference estimation in bifactor ordered-categorical data. In the No Differential Item Functioning (DIF) conditions, the data generation conditions varied were sample size, the number of categories per item, effect size of the general factor mean difference, and the size of specific factor loadings; in data analysis, misspecification conditions were introduced in which the generated bifactor data were fit using a unidimensional model, and/or ordered-categorical data were treated as continuous data. In the DIF conditions, the data generation conditions varied were sample size, the number of categories per item, effect size of latent mean difference for the general factor, the type of item parameters that had DIF, and the magnitude of DIF; the data analysis conditions varied in whether or not setting equality constraints on the noninvariant item parameters.

Results showed that falsely fitting bifactor data using unidimensional models or failing to account for DIF in item parameters resulted in estimation bias in the general factor mean difference, while treating ordinal data as continuous had little influence on the estimation bias as long as there was no severe model misspecification. The extent of estimation bias produced by misspecification of bifactor datasets with unidimensional models was mainly determined by the degree of unidimensionality (i.e., size of specific factor loadings) and the general factor mean difference size. When the DIF was present, the estimation accuracy of the general factor mean difference was completely robust to ignoring noninvariance in specific factor loadings while it was very sensitive to failing to account for DIF in threshold parameters. With respect to ignoring the DIF in general factor loadings, the estimation bias of the general factor mean difference was substantial when

the DIF was -0.15, and it can be negligible for smaller sizes of DIF. Despite the impact of model misspecification on estimation accuracy, the power to detect the general factor mean difference was mainly influenced by the sample size and effect size. Serious Type I error rate inflation only occurred when the DIF was present in threshold parameters.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere appreciation to my advisor Dr. Marilyn Thompson for the guidance on my dissertation and continuous support of my Ph.D study. As a mentor, Dr. Thompson does not only provide help in my academic development, but also cares about my personal life. I feel very lucky to have such a wonderful mentor.

Also, I would like to thank other committee members for my dissertation: Dr. Roy Levy and Dr. Holly O'Rourke, who provide me much really helpful advice and inspire me to view my research questions from new angles.

Then, I would like to thank my parents. They taught me how to keep calm whatever happens and how to be a person with inner confidence. Although they were not living together with me in the past nine years, they would give me support whenever I need.

Finally, my thanks goes to my husband and my children. My husband is very intelligent and he is my soul mate. He supports me spiritually and makes me stronger. My children are my sweet burden. I love them and they make me happy.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1: Introduction

Overview

The bifactor model is widely used in measurement models when sets of items are grouped into clusters. For example, in a reading comprehension test, a cluster may be formed when a set of questions are given based on a specific reading passage. These clusters can be considered as testlets, which are very common in both cognitive and non-cognitive tests (e.g., Chen, West, & Sousa, 2006; Gignac & Watkins, 2013; Min & He, 2014; Reise, Morizot, & Hays, 2007). Testlet-based items are desirable mainly in the following two circumstances. First, the construct (e.g., the depression construct) to be measured may consist of several related facets (e.g., negative mood, social withdrawal, poor cognitive functioning, etc.). Second, as in the example mentioned earlier, context-dependent items may be based on a common stimulus (e.g., a reading passage). In these circumstances, the bifactor model can be an appropriate representation of the construct when an assessment is designed to measure a strong common trait despite the existence of testlets (Reise, 2012). In the bifactor model, a general factor is hypothesized to underlie all items, and each item is specified to load on at most one of the specific factors, which explains the additional common variance among a set of items beyond the influence of the general factor. It is assumed that the general factor and all specific factors are orthogonal with each other.

The bifactor model was initially applied as a special case of the confirmatory factor analysis (CFA) model for continuous items (Holzinger & Swineford, 1937). To accommodate a wider range of measurement applications, the bifactor model was extended for use with binary data within the item-response theory (IRT) framework by Gibbons and

Hedeker in 1992. Gibbons et al. (2007) introduced the bifactor IRT model for polytomous data.

The single-group bifactor model based on ordered-categorical data has been widely applied and studied in recent years. In both applied research and simulation studies (e.g., DeMars, 2006; Immekus & Imbrie, 2008; Min & He, 2014; Reise et al., 2007; Rijmen, 2010), researchers are mostly interested in comparing the bifactor IRT model with other competitive IRT models for items which may form testlets, including unidimensional models (only the common trait is modeled), testlet models (constraints are placed on the relationship between the general factor loadings and the specific factor loadings of the bifactor model), second-order IRT models (equivalent with a testlet model in which a proportional constraint is specified between the general factor loadings and the specific factor loadings for items within each testlet), and correlated-factors models (only the specific factors are modeled and the specific factors can be correlated with each other). These models can be compared based on parametric methods by utilizing exploratory and confirmatory models. Nonparametric DIMTEST (Stout, Douglas, Junker, & Roussos, 1999) is used in some research (e.g., DeMars, 2006) to explore essential unidimensionality. If a unidimensional model is deemed adequate, there would be no need for subsequent comparisons of varied multidimensional IRT models.

Given that the unidimensional model, the testlet model, the second-order model and the correlated-factors model are all nested within the bifactor model, the bifactor model plays an important role in determining dimensionality issues for testlet-based items. As suggested by DeMars (2013), another important utility of the bifactor model is that more meaningful general factor scores can be obtained after accounting for the specific factors.

Also, bifactor models can be used to estimate the extent to which a subset of items can discriminate the ability reflected by the subdomain after the common variance due to the general factor is partialed out such that a decision can be made regarding the utility of forming subscale scores (Reise et al., 2007). In addition, the unique contribution of each specific factor (or general factor) to prediction of an external variable after controlling for the general factor (or specific factors) can be estimated using a bifactor model (Chen et al., 2006).

With respect to estimation, the bifactor model with ordered-categorical data can be estimated by both full-information estimation (e.g., marginal maximum likelihood; Bock & Aitkin, 1981) under the IRT framework (e.g., Gibbons & Hedeker, 1992; Gibbons et al., 2007) and limited-information estimation (e.g., weighted least squares) within the framework of structural equation modeling (SEM; Reise, 2012). Unlike full-information estimation in which the entire response vector of each test taker is utilized for computation, the limited-information estimator is implemented based on tetrachoric or polychoric correlations among items. It has been shown that the two-parameter normal-ogive IRT model is equivalent to the factor analytic model for ordinal categorical data (Kamata & Bauer, 2008; Takane & de Leeuw, 1987).

In the field of consumer research, organizational research, and clinical studies, researchers are frequently interested in latent mean differences across different populations in terms of demographic characteristics, cultures, and backgrounds. In addition to the utilities mentioned above, with bifactor models, latent mean differences in both the general factor and the specific factors can be estimated across groups (Chen et al., 2006). As suggested by Schmitt and Kuljanin (2008), the establishment of measurement invariance

is crucial for the comparisons of latent means or other structural coefficients because these subsequent analyses might be meaningless if directly assuming measurement invariance. Measurement invariance holds if a measuring device works in the same way across varied conditions (i.e., different populations, different time points) that are irrelevant to the attribute being measured (Millsap, 2011). Differential item functioning (DIF) is considered as a between-group difference between item parameters, or item response functions given the same score on the latent continuum, that determine the item response function in different groups. Under the IRT framework, there are multiple methods for DIF detection in item parameters, and the most commonly applied method is likelihood ratio (LR) tests. In addition, the DIF can be detected under the CFA framework as well using traditional multiple-group CFA models, categorical multiple-group CFA models and multiple-indicator-multiple-causes (MIMIC) models. It has been consistently agreed that latent mean comparisons can be conducted under conditions of partial invariance (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). However, there are no consistent opinions about the extent to which partial invariance is allowed without compromising estimation accuracy and power for tests of latent mean differences. Only a few simulation studies have focused on the factors influencing latent mean comparisons under both IRT and CFA frameworks (e.g., De Beuckelaer & Swinnen, 2018; Jones & Gallo, 2002). Their results indicated that one of the major factors that resulted in bias in latent mean difference estimation was failing to account for DIF.

For the multiple-group data with unknown structure, dimensionality issues need to be explored first before testing DIF and estimating between-group latent mean differences. The consequences of fitting bifactor data with unidimensional models have been studied

in some single-group studies (e.g., DeMars, 2006). To the best of my knowledge, only one study focused on this issue for multiple-group data (Fukuhara & Kamata, 2011), and as shown in their results, for the generated bifactor binary response data, DIF can be better detected using the bifactor IRT model in comparison with the unidimensional model.

Although selecting the appropriate model and correctly detecting DIF in the item parameters are the prerequisite for estimating latent mean differences, they might not be achieved in reality. Thus, the main purpose of this study is to explore the robustness of latent mean comparisons for the general factor underlying bifactor, ordered-categorical data to misspecification of the dimensionality of data structure and the equality constraints on noninvariant item parameters under varied research conditions. In this chapter, I first introduce the bifactor IRT model in terms of its specification, estimation, applications, and important utilities. Next, I focus on different methods for DIF detection and latent mean comparisons. Finally, I illustrate factors that influence DIF detection procedures and latent mean comparisons based on findings of previous simulation studies. Following this review, the proposed simulation study is presented in the methods chapter. It is expected that the results of this study will provide recommendations for researchers who are interested in the latent mean difference of the general factor despite the existence of the specific factors in bifactor, ordered-categorical data.

Introduction to Bifactor IRT Models

The following section starts with an introduction to basic principles of IRT models based on unidimensional IRT models, which can be extended to multidimensional IRT models. Then the dimensionality issues are addressed. To be specific, the general forms of multidimensional models are illustrated first. Then I introduce the bifactor IRT model, a

hierarchical multidimensional IRT model, in terms of its specification, estimation, applications, and important utilities. After that, I discuss the role of bifactor models in exploring the dimensionality issues of IRT models by comparing the bifactor model with its competing alternative models.

Basic Principles of Item Response Theory Models

Item response theory was introduced around 1950s as a relatively recent alternative to classical test theory (CTT). Unlike CTT that focuses on total observed scores, IRT focuses on each item. IRT places the person characteristics and item characteristics on the same latent continuum, and the item response function (IRF) specifies the function that relates the probability of responses to both person characteristics and item characteristics (de Ayala, 2009).

For dichotomous data, the common IRT models are the one-parameter logistic (1PL) model, the two-parameter logistic (2PL) model, and the three-parameter logistic (3PL) model. The 2PL model was the focus of the current study. The IRF of the 2PL model is shown below:

$$P(x_{ij} = 1|\theta_i, a_j, b_j) = \frac{e^{a_j(\theta_i - b_j)}}{1+e^{a_j(\theta_i - b_j)}} \quad (1)$$

where x_{ij} denotes person *i*'s response to item *j*, θ_i is the latent ability parameter for person *i*, a_j is the discrimination parameter for item *j*, b_j is the difficulty parameter for item *j*, and $P(x_{ij} = 1|\theta_i, a_j, b_j)$ denotes the probability of person *i* correctly answering item *j*. To obtain a unique solution for the discrimination parameter for item *j* (a_j), a common identification method is to set the variance of the latent ability (θ) distribution to 1; to obtain a unique solution for the difficulty parameter for item *j* (b_j), a common identification method is to set the mean of the latent ability (θ) distribution to 0.

For 2PL models, each item has its own difficulty parameter that represents the point on the latent ability scale where the probability of passing this item is .50. The discrimination parameter reflects the steepness of an item. For an item with a larger value of the discrimination parameter, the probability of passing it changes more quickly than another item with relatively smaller discrimination parameter around the neighborhood of their corresponding difficulty parameters. In the 2PL model, the discrimination parameters vary across items.

For polytomous data, the common IRT models include the partial credit model (PCM), the rating scale model (RSM), the generalized partial credit model (GPCM) and the graded response model (GRM). The GRM was focused in the current study. The GRM has the same specification as the 2PL model. The equation for the GRM is:

$$P(x_{ij} \geq x|\theta_i, a_j, b_{jx}) = \frac{e^{a_j(\theta_i - b_{jx})}}{1+e^{a_j(\theta_i - b_{jx})}} \quad (2)$$

where b_{jx} denotes the threshold parameter representing the point on the latent ability scale where the probability that a response above *x* is .50, and $P(x_{ij} \geq x|\theta_i, a_j, b_{jx})$ is the probability for person *i* to give a response above *x*. Note that for an item with *C* categories there are *C* -1 threshold parameters. $P(x_{ij} \geq x|\theta_i, a_j, b_{jx})$ equals 1 when x is 0.

Based on Equation 3, the probability of giving a response of *x* can be obtained from:

$$P(x_{ij} = x|\theta_i, a_j, b_{jx}) = P(x_{ij} \geq x|\theta_i, a_j, b_{jx}) - P(x_{ij} \geq x+1|\theta_i, a_j, b_{j(x+1)}) \quad (3)$$

For IRT models with polytomous data, researchers can get an overall picture of the probabilities for an item using the expected score function. The expected score is calculated by summing the products of the number assigned to each category and probability of this category given the latent ability, so the expected score function describes the relationship

between the latent ability and the expected item score. When there are only two categories, the expected score function is actually the item response function.

There are three fundamental assumptions underlying commonly applied IRT models (de Ayala, 2009). First, it is assumed that only one latent variable determines the probability of observed responses (unidimensionality assumption). Second, it is assumed that the item responses are uncorrelated with each other after controlling for the latent variable (local independence assumption). Third, it is assumed that the IRT model follows a specific form specified by the model (functional form assumption).

The most commonly applied estimation method under IRT framework is marginal maximum likelihood (MML; Bock & Aitkin, 1981). In MML, item parameters are estimated first using the marginal distribution in which person parameters are removed from the marginalization process. After obtaining item parameters, person parameters can be estimated using either maximum likelihood estimation (MLE) or the Bayesian method (de Ayala, 2009). The drawback of MLE is that it cannot estimate latent scores for examinees with zero correct answers or perfect scores. Expected a posterior (EAP) and maximum a posterior (MAP) are two specific strategies for the Bayesian method. In addition to MML, Bayesian estimation with Markov Chain Monte Carlo (MCMC) has gained popularity in recent years.

The 2PL IRT model and the GRM can be estimated within the framework of the CFA model because of their equivalency with categorical CFA models (Wirth & Edwards, 2007). In categorical CFA models, it is assumed that continuous latent response variates underlie the ordered-categorical data. Using x_j to represent the observed discrete response

variable for item j and x_j^* to represent the underlying latent response variate for x_j, one could express the relationship between x_j and x_j^* as:

$$x_j = 0, if\ x_j^* < \tau_{j1}$$

$$x_j = x, if\ \tau_{jx} < x_j^* < \tau_{j(x+1)}$$

$$x_j = C - 1, if\ \tau_{j(C-1)} < x_j^* \quad (4)$$

where τ_{jx} is the *x*th threshold parameter for item *j*. Using the one-factor CFA model as an example, the equation relating the common latent factor to the latent response variates is:

$$X^* = \mathbf{\Lambda_x^*}\xi^* + \mathbf{\delta}^* \quad (5)$$

where X^* is the vector containing latent response variates, $\mathbf{\Lambda_x^*}$ is the loading vector, ξ^* denotes the common latent factor, and $\mathbf{\delta}^*$ is the residual vector. The model implied variance and covariance matrix $\mathbf{\Sigma}^*$can be expressed as:

$$\mathbf{\Sigma}^* = \mathbf{\Lambda_x^*}\Phi^*\mathbf{\Lambda_x^{*'}} + \mathbf{\Theta_\delta^*} \quad (6)$$

where Φ^* is the variance of the common latent factor, $\mathbf{\Lambda_x^{*'}}$ is the transpose vector of $\mathbf{\Lambda_x^*}$, and $\mathbf{\Theta_\delta^*}$ is the variance and covariance matrix for the residuals. In order to identify this model, the mean of the common factor is fixed to 0, the variance of the common latent factor is fixed to 1, and the variances of latent variates are fixed to 1. Categorical CFA models are estimated using tetrachoric or polychoric correlations among the items which can be considered as the estimates of Pearson correlations among the latent response variates. Like the CFA models with continuous data, parameters in the categorical CFA models are estimated to minimize the differences between the model-implied variance-covariance matrix and the data variance-covariance matrix. Given that only the tetrachoric or polychoric correlations estimated based on proportion of responses in the observed

contingency table are used as data input, estimation methods applied for CFA models with ordered-categorical data are called limited-information analysis, which is named in contrast to full-information analysis (e.g., MML) that utilizes all information of the data. The WLSMV estimator of Mplus (Muthén & Muthén, 2010) is commonly applied for estimating model parameters for categorical CFA models.

In unidimensional models, the loading parameters and threshold parameters obtained in categorical CFA models can be converted to discrimination parameters and difficulty (or threshold) parameters for the corresponding equivalent 2PL (or the GRM) IRT model using the following formulas:

$$a_j = \frac{1.7\lambda_j^*}{\sqrt{1-\lambda_j^{*2}}}$$

$$b_j(b_{jx}) = \frac{\tau_{j(x)}}{\lambda_j^*} \quad (7)$$

where λ_j^* and τ_{jx} denote the standardized factor loading and threshold parameter for item j in the categorical CFA model, and a_j and $b_j(b_{jx})$ represent the discrimination parameter and difficulty parameter (threshold parameter).

For both IRT models and categorical CFA models, global model fit and local model fit indices can be obtained using the commonly applied software (e.g., IRTPRO for IRT models, Cai, Thissen & du Toit, 2011; Mplus for CFA models, Muthén & Muthén, 2010). Nested models can be compared using the likelihood ratio (LR) test for IRT models and the chi-square difference test with corrections for categorical CFA models (i.e., DIFFTEST option of the Mplus).

Although IRT models can be estimated within the framework of CFA for some IRT forms (i.e., 2PL model and the GRM), the IRT framework can provide some unique

features. For example, the lack of local fit for a 2PL IRT model might suggest the existence of a pseudo-guessing parameter. Another important feature of IRT models is that information functions are obtained for each item and the test. The information function describes how well an item (or a test) can discriminate among examinees with different latent ability scores. For unidimensional IRT models, the amount of information an item or a test can provide for an examinee depends on this person's latent ability level. The discrimination parameter determines the maximum information an item can provide. The test information function is the sum of all the item information functions such that the length of a test also determines the information of this test. When estimating latent score ability using MLE, the standard error of an estimate is the inverse square root of the test information given this person's estimated ability level.

Reise (2012) suggested two potential problems regarding the equivalence between the IRT model and CFA model. First, the interpretation of loading parameters in the CFA model might differ from the converted discrimination parameters in the IRT model in terms of the magnitude. Second, it might not be appropriate to use the model fit obtained from a linear CFA model to interpret the model fit for a non-linear IRT model because these models are estimated based on different assumptions.

As an alternative to CTT based on the true score model, IRT provides several unique utilities (Reise & Henson, 2003). First, in IRT, an individual's location on the latent continuum is estimated and each item can have an unequal contribution in estimating latent ability scores. Second, item characteristics and person characteristics are independent with each other in IRT which is the foundation of computerized adaptive testing (CAT; Wainer, 2000) and IRT based linking methods. Third, the unidimensional assumption and the local

independence assumption of IRT are often tested in real data and the dimensionality issues are discussed in more detail later. Fourth, in order to have comparable test scores across different groups of people, one needs to assess whether or not items function the same way across these groups. IRT provides systematic methods for detecting differential item functioning (DIF).

The General Form of Multidimensional IRT Models

As suggested by Reise et al. (2007), measures differ in their degree of conceptual breadth. A measure is considered to be broad if it contains relatively heterogeneous items and it is considered to be narrow if it contains relatively homogeneous items. For example, a measure of depression might be considered as a broad measure because it contains multiple aspects of the depression construct such as negative mood, social withdrawal, poor cognitive functioning, somatic concerns, and suicidal ideation. In contrast, if a test is designed to measure somatic concerns, it is considered to be relatively narrow. For a relatively narrow measure, it is more likely to specify a unidimensional model; for a relatively broad measure, it is more likely to explore the dimensionality issue. Another circumstance in which multidimensional IRT models are desirable is when the items of a test are indicators of more than one skill (Ackerman, 2005). In fact, all the assessments measure multiple dimensions, and whether examinees vary on those dimensions which items strongly load on determines the dimensionality of the model. For example, scores on mathematics problem solving items may reflect both mathematics skills and reading skills, so a multidimensional IRT model is more desirable. However, if the test takers only differ in one of the skills, a unidimensional model is preferred (Ackerman, 2005). In multiple-

group models, whether distributions of different groups vary on a given strongly related dimension also determines the dimensionality of the model.

The multidimensional IRT models have two types of structures which are between-item multidimensionality and within-item multidimensionality (Adams, Wilson & Wang, 1997). For the model with between-item multidimensionality, each item discriminates on only one of the several dimensions and these dimensions might be correlated with each other, which corresponds to simple structure in factor analysis models. For the model with within-item multidimensionality, some of the items discriminate on more than one dimension, which corresponds to complex structure in factor analysis models.

For the items that discriminate more than one dimension, either compensatory models (Reckase, 1985) or non-compensatory models (also called the partial compensatory model; Sympson, 1978) can be applied. Taking an example of the 2PL IRT model, the equation for the compensatory model is:

$$P(x_{ij} = 1|\boldsymbol{\theta_i}, \boldsymbol{a_j}, d_j) = \frac{e^{(\boldsymbol{a_j}\boldsymbol{\theta_i}' + d_j)}}{1+e^{(\boldsymbol{a_j}\boldsymbol{\theta_i}' + d_j)}} \quad (8)$$

where $\boldsymbol{\theta_i}$ is the 1 x m vector containing multiple latent scores for person i, $\boldsymbol{a_j}$ is the 1 x m vector containing discrimination parameters of item j with respect to corresponding latent abilities, and d_j is the intercept parameter of item j. Although each latent ability has a corresponding discrimination parameter, only one intercept parameter is estimated because the difficulty parameters with respect to multiple latent abilities are indeterminate. In the compensatory model, the low ability of an examinee on one dimension can be compensated by the high ability of this examinee on another dimension in terms of the probability of passing an item. In contrast to the compensatory model in which multiple latent abilities

are added together in the logit, the non-compensatory model is specified as the product of multiple unidimensional models. Using the example of the 2PL model and supposing that there are two underlying abilities, one can express the equation of the non-compensatory model as:

$$P(x_{ij} = 1|\theta_{i1}, \theta_{i2}, a_{j1}, a_{j2}, b_{j1}, b_{j2}) = \frac{e^{a_{j1}(\theta_{i1} - b_{j1})}}{1+e^{a_{j1}(\theta_{i1} - b_{j1})}} \text{ x } \frac{e^{a_{j2}(\theta_{i2} - b_{j2})}}{1+e^{a_{j2}(\theta_{i2} - b_{j2})}} \qquad (9)$$

where θ_{i1}, θ_{i2} are latent abilities for person i, a_{j1}, a_{j2} are corresponding discrimination parameters, and b_{j1}, b_{j2} are corresponding difficulty parameters. In the non-compensatory model, even if a person has very high ability in one dimension and extremely low on another dimension, the probability for this person to pass an item is still very low. As shown in the research of Babcock (2011), non-compensatory models can be estimated using Bayesian methods.

An Introduction to Bifactor IRT Models

Bifactor IRT models are hierarchical multidimensional models in which a general factor explains the common variance among all the items and specific factors are modeled to explain the common variance independent of the general factor (Reise, 2012). Each item is allowed to load on at most one of the specific factors. It is assumed that the general factor is orthogonal with the specific factors and there are no correlations among the specific factors. Although the orthogonality assumption might be hard to achieve for real data, Reise (2012) suggested that the specific factors cannot be considered as residualized factors that explain the additional common variance beyond the general factor if they are allowed to correlate with the general factor. Correlations among the specific factors may indicate the existence of other factor(s) that complicate the structure of the data. Despite the importance of orthogonality assumption, Jeon, Rijmen, and Rabe-Hesketh (2011)

suggested that relaxing this assumption in multiple-group bifactor models when it is violated in one of the groups could improve estimation accuracy for the DIF.

The bifactor model was introduced by Holzinger and Swineford (1937) based on the factor analysis model. Gibbons and Hedeker (1992) used the EM algorithm for marginal maximum likelihood estimation to analyze binary data under the framework of IRT. In 2007, Gibbons et al. (2007) applied bifactor IRT models for polytomous data. As pointed out by Reise (2012), the bifactor model has become an important representation of multidimensional structure and has gained increasing popularity in research and applications for both IRT and SEM in recent years.

In the bifactor IRT model, the general factor represents a broader concept (e.g., depression) or the main trait intended to measure (e.g., mathematics skills for mathematics problem solving items) whereas the specific factors represent narrower concepts (e.g., negative mood, social withdrawal, poor cognitive functioning, somatic concerns, and suicidal ideation) or the trait not intended to measure (e.g., reading skills for mathematics problem solving items). Although researchers are primarily interested in individual differences in the general factor, clusters of items are designed for the following reasons. First, the majority of psychological constructs are complex constructs including multiple facets, such that subdomains of items are needed to improve content validity (Reise, 2012). Second, in cognitive tests, context-dependent items are desirable for measuring higher-level abilities such as problem-solving skills, and the application of common stimulus (e.g., a reading passage) provides a good way to save examinees' time (DeMars, 2006).

The bifactor structure can be applied to both compensatory IRT models and non-compensatory IRT models although almost all the research and applications focused on

compensatory bifactor models (Desa, 2012). The general form of multidimensional item response function shown in Equations 8 and 9 can be applied to bifactor models. Compensatory IRT models were focused in the current study. Taking the example of Equation 8 and supposing that item *j* loads on the *k*th specific factor, the item response function of the bifactor model can be written as:

$$P(x_{ij} = 1|\theta_{iGEN}, \theta_{iGRk}, a_{jGEN}, a_{jGRk}, d_j) = \frac{e^{(a_{jGEN}\theta_{iGEN} + a_{jGRk}\theta_{iGRk} + d_j)}}{1+e^{(a_{jGEN}\theta_{iGEN} + a_{jGRk}\theta_{iGRk} + d_j)}} \quad (10)$$

where θ_{iGEN} denotes the general factor score for person *i*, θ_{iGRk} denotes the specific factor score for person i, a_{jGEN} is the discrimination parameter of item *j* for the general factor, a_{jGRk} is the discrimination parameter of item *j* for the *k*th specific factor, and d_j represents the item intercept of item *j* which is the log-odds of correct responses when θ_{iGEN} and θ_{iGRk} are all zero. The discrimination parameter in the bifactor model reflects how well an item can discriminate examinees along with a given dimension (general dimension or specific dimension) of the item response surface. A multidimensional information surface is used to indicate the information provided by an item for each point on the ability plane and it is formed for each ability composite (direction on the ability plane). To compare the degree of difficulty cross items based on a bifactor model, a multidimensional difficulty (MDIFF) parameter can be calculated as $-d_j / \sqrt{{a_{jGEN}}^2 + {a_{jGRk}}^2}$. Item with higher MDIFF is considered to be more difficult, whereas the item with lower MDIFF is considered to be easier.

With respect to person parameters, it is assumed that in the bifactor model, the general factor scores and specific factor scores are from a multivariate normal distribution with orthogonal dimensions (Gibbons & Hedeker, 1992). In most of the cases, researchers are mostly interested in individual differences on the general factor when applying a

bifactor model, which are reflected by θ_{iGEN}. Sometimes researchers also want to evaluate an examinee's performance on the subscale. It should be noted that specific factor score estimates cannot be directly used for scaling individual differences on the subscale because they reflect the residualized factor scores beyond the information provided by the general factor (DeMars, 2013). But they can be used to evaluate examinees' strengths and weaknesses on the subscale after controlling for the general factor. To estimate an examinee's overall performance on a given subscale, one needs to either use the correlated-factors model or some relatively sophisticated methods based on bifactor models such as the composite score of both the general factor score and the residualized factor score (DeMars, 2013) and the restricted bifactor model (Chang, 2015).

To identify a bifactor IRT model with freely estimated discrimination parameters for both the general factor and specific factors, the means of the general factor and specific factors need to be fixed to 0, and the variances of the general factor and specific factors need to be fixed to 1. In some special cases of bifactor models, some more constraints might be needed for identification purpose. For example, if a specific factor has only two indicators, equality constraints need to be placed on the item discrimination parameters for these two indicators.

The EM algorithm for MML is commonly used for estimation of bifactor IRT models (Gibbons & Hedeker, 1992; Gibbons et al., 2007). In multidimensional IRT models, the likelihood function of responses of *N* persons for *p* binary items can be written as:

$$L\,(X \mid \boldsymbol{\Gamma}, \boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{j=1}^{p} P\left(x_{ij} = 1 \middle| \Gamma_j, \boldsymbol{\theta_i}\right)^{x_{ij}} P\left(x_{ij} = 0 \middle| \Gamma_j, \boldsymbol{\theta_i}\right)^{1-x_{ij}} \qquad (11)$$

where X represents the responses of N persons for p items, $\boldsymbol{\Gamma}$ contains all the item parameters, $\boldsymbol{\theta}$ is the latent scores of all the dimensions for all the persons, x_{ij} represents

the response of item j for person i, Γ_j contains the item parameters for item j (e.g., discrimination parameters for all the dimensions, item intercept), and $\boldsymbol{\theta_i}$ contains latent scores for all the dimensions for person i. In the MML, people are considered to be randomly drawn from a multidimensional distribution $g(\boldsymbol{\theta})$. Supposing that there are k dimensions of latent abilities $(\theta_1, \theta_2, \ldots, \theta_k)$ underlying the responses, the marginal likelihood function of the responses can be written as:

$$L(X \mid \boldsymbol{\Gamma}) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} L(X \mid \boldsymbol{\Gamma}, \boldsymbol{\theta})\, g(\boldsymbol{\theta}) d\theta_1 d\theta_2 \ldots d\theta_k \quad (12)$$

For the bifactor IRT models, all the items load on one general factor and each item loads on at most one of the specific factors. Thus, the likelihood function only needs to integrate over two dimensions regardless of the number of total dimensions involved which greatly simplifies the integration process for multidimensional IRT models (Gibbons et al., 2007). Once item parameters are obtained and the model fit is acceptable, latent scores of the general factor and residualized factors can be estimated using MLE or Bayesian methods (EAP or MAP).

Bifactor models with ordered-categorical data can be estimated within the CFA framework (Reise, 2012). In bifactor models, the relationship between the observed ordinal variables and their corresponding continuous latent variates follows the same rule of unidimensional models (shown in Equation 4), and the equation relating the general factor and specific factors to the latent response variates is

$$X^* = \boldsymbol{\Lambda}_{\mathbf{x}}^* \boldsymbol{\xi}^* + \boldsymbol{\delta}^* \quad (13)$$

where X^* is the vector containing latent response variates, $\boldsymbol{\Lambda}_{\mathbf{x}}^*$ is the loading matrix, $\boldsymbol{\xi}^*$ denotes the vector containing the general factor and specific factors, and $\boldsymbol{\delta}^*$ is the residual vector. For example, if a bifactor model has 9 items and three specific factors, items 1-3

load on the first specific factor, items 4-6 load on the second specific factor, and items 7-9 load on the third specific factor, then the vector of $\boldsymbol{\xi}^*$ is:

$$\boldsymbol{\xi}^* = \begin{bmatrix} \xi^*_{GEN1} \\ \xi^*_{GR1} \\ \xi^*_{GR2} \\ \xi^*_{GR3} \end{bmatrix} \tag{14}$$

where ξ^*_{GEN1} is the general factor, ξ^*_{GR1} is the first specific factor, $\xi^*_{GR2}$is the second specific factor, and ξ^*_{GR3} is the third specific factor. The loading matrix $\boldsymbol{\Lambda}^*_{\mathbf{x}}$ has the following pattern:

$$\boldsymbol{\Lambda}^*_{\mathbf{x}} = \begin{bmatrix} \lambda^*_{GEN1,1} & \lambda^*_{GR1,1} & 0 & 0 \\ \lambda^*_{GRN2,1} & \lambda^*_{GR2,1} & 0 & 0 \\ \lambda^*_{GEN3,1} & \lambda^*_{GR3,1} & 0 & 0 \\ \lambda^*_{GEN4,1} & 0 & \lambda^*_{GR4,2} & 0 \\ \lambda^*_{GEN5,1} & 0 & \lambda^*_{GR5,2} & 0 \\ \lambda^*_{GEN6,1} & 0 & \lambda^*_{GR6,2} & 0 \\ \lambda^*_{GEN7,1} & 0 & 0 & \lambda^*_{GR7,3} \\ \lambda^*_{GEN8,1} & 0 & 0 & \lambda^*_{GR8,3} \\ \lambda^*_{GR9,1} & 0 & 0 & \lambda^*_{GR9,3} \end{bmatrix} \tag{15}$$

where λ^*_{GEN} denotes the factor loadings relating the general factor to the latent response variates, and λ^*_{GR} denotes the factor loadings relating the specific factors to the latent response variates. The limited-information estimation method based on tetrachoric or polychoric correlations are used for bifactor models with ordered-categorical data. In contrast to the full-information estimation in which the entire response vectors are made use of when estimating item parameters, in the limited-information estimation, only the observed response contingency table among the items are used for estimating model parameters.

In multidimensional models, The loading parameters and threshold parameters obtained in categorical CFA models can be converted to discrimination parameters and

item intercept parameters for the corresponding equivalent 2PL (or the GRM) IRT model using the following formulas:

$$a_{jp} = \frac{1.7\lambda_{jp}^{*}}{\sqrt{1 - \sum_{p=1}^{P} \lambda_{jp}^{*2}}}$$

$$d_j(d_{jx}) = \frac{-1.7\tau_{j(x)}}{\sqrt{1-\sum_{p=1}^{P} \lambda_{jp}^{*2}}} \quad (16)$$

where λ_{jp}^{*} denotes the standardized factor loading of dimension *p* for item *j* and τ_{jx} denotes the threshold parameter for item j in the categorical CFA model, and a_{jp} and $d_j(d_{jx})$ represent the discrimination parameter of dimension *p* and item intercept parameter.

In most applications of bifactor models, researchers are more interested in individual differences on the general factor. As suggested by DeMars (2013), employing bifactor IRT models, researchers can get pure estimates of the common latent trait because the common variances due to the specific factors are accounted for. Reise et al. (2007) compared multiple models using the data from the Consumer Assessment of Healthcare Providers and Systems. They found that the discrimination parameters on the general factor increased for some of the items after modeling the specific factors using a bifactor model in comparison with the corresponding parameters obtained from a unidimensional model. Thus, they argued that these items became more meaningful measures after controlling for the specific factors.

Researchers are also interested in examinees' performance on subscales sometimes. Although an examinee's subscale score cannot be directly estimated using a bifactor model, his or her strengths and weaknesses on each subscale beyond the influence of the general factor are reflected by the residulized factor score (DeMars, 2013), so researchers can

employ a bifactor model to determine whether it is meaningful to report scores for a subscale (Reise et al, 2007). In the study of Reise et al. (2007), although the item loadings on each subscale were fairly strong in the correlated-factors model, when using the bifactor model, most of the items had larger discrimination parameters on the general factor than those on the specific factors. For example, one of the items had a discrimination parameter of 1.30 in the correlated-factors model. If we only looked at this result, we might think that it is a good measure of the subdomain. However, when employing the bifactor model, this item had a discrimination parameter of 1.27 on the general factor and 0.38 on the specific factor, which suggested that the unique contribution of the subdomain on this item was very small after controlling for the general factor. DeMars (2013) summarized that subscale scores are useful when items have high discrimination parameters on the subscale(s) in the bifactor model whereas scoring the subscale(s) might be redundant if the item discrimination parameters of the subscale(s) are very low.

Reise (2012) proposed another method to determine whether subscale scores should be formed using an index omega subscale (ω_s). ω_s can be used to indicate the model-based reliability for a subscale after controlling for the general factor in a bifactor model. Using subscript *m* to denote the items loading on the *k*th subscale, omega subscale for the *k*th subscale (ω_{sk}) can be calculated using the following formula:

$$\omega_{sk} = \frac{(\sum \lambda_{mGRk})^2}{(\sum \lambda_{mGEN})^2 + (\sum \lambda_{mGRk})^2 + \sum \theta_m} \quad (17)$$

where λ_{mGEN} represents the standardized general factor loadings for the items relating to the *k*th subscale, λ_{mGRk} represents the standardized specific factor loadings for these items, and θ_m represents the error variances for these items. Note that all of these parameters are obtained under a CFA framework. Omega subscale shown in Equation 17 reflects the

proportion of common variance among a subset of items due to the specific factor beyond the influence of the general factor. In the study of Reise (2012), five specific factors were modeled in a bifactor model, and the indices of omega subscale were shown to be .21, .32, .26, .44 and .22, respectively. The proportion of common variance of a subset of items that was due to both the general factor and their corresponding specific factor was also calculated in Reise's study (2012), and they were .62, .66, .67, .62, and .66, respectively. Based on these results, Reise (2012) pointed out that the reliable variance on the subscales was little if the variance due to the general factor was partialed out, so he concluded that there was no need to report subscale scores if total scores were given.

Reise, Moore, and Haviland (2010) pointed out a problem of reporting subscale scores in predicting external variables. They argued that the multicollinearity among the subscales might make it harder to precisely estimate the unique effect of each subdomain on outcome variables. In this case, bifactor models would be desirable in which the unique contribution of the general factor and each subdomain to the external outcome variables can be estimated (Chen et al., 2006; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012; Gustafsson & Balke, 1993).

Given that bifactor models are appropriate representations of many complex constructs measured in both cognitive and non-cognitive tests, bifactor models can offer utilities in the areas where an IRT model is desirable. For example, testlet-based items are commonly applied in CAT (Wainer, Bradlow, & Du, 2000). In cognitive tests, context-based items are often desirable in measuring some higher-level skills (DeMars, 2006). Also, making utility of context-based items can improve test efficiency because it would be time consuming if an examinee only needs to answer one question after reading a long passage.

In non-cognitive CAT, the construct to be measured might include multiple aspects (e.g., Haley et al., 2009). In these circumstances, bifactor modeling can help researchers retain their primary interest in the general factor while avoiding violation of local independency assumption due to forcing all the items onto only one dimension.

Another important utility of IRT models is to help researchers link scales across multiple measures such that test scores from different measures can be comparable (Reise & Henson, 2003). As suggested by Reise et al. (2007), linking scales based on bifactor models are usually more complicated, but if researchers are only interested in linking measures onto the general factor, standard linking methods can be used. Li (2011) applied the bifactor model for vertical scaling in which measures with similar construct but different difficulty levels were linked onto the same scale such that the test scores of students from different grades can be comparable and the growth of a given student can be tracked. In Li's study (2011), the general factor was used to represent the common vertical scale across grades and the specific factors were used to represent the shifted construct specific to each grade.

Comparisons of Bifactor IRT Models with Competing Models

The applications of bifactor IRT models in exploring dimensionality issues are of great importance to researchers. The common alternative models to bifactor models include the unidimensional model, the testlet-effects model, the second-order model and the correlated-factors model (e.g., DeMars, 2006; Immekus & Imbrie, 2008; Min & He, 2014; Reise et al., 2007; Rijmen, 2010), and all these models are nested within the bifactor model.

Both confirmatory models and exploratory models can be used for comparisons among these competitive models (e.g., Reise et al., 2007; Reise et al., 2010). When

applying confirmatory models, the unidimensional model, the testlet-based model, the second-order model and the correlated-factors model can be compared with the bifactor model using either the chi-square different test with corrections (i.e., DIFFTEST option of Mplus) under the CFA framework (e.g., Reise, 2012) or the LR test under the IRT framework (e.g., DeMars, 2006; Immekus & Imbrie, 2008). When implementing exploratory models for potentially multidimensional data, researchers can conduct both the standard exploratory factor analysis (e.g., exploratory principle axis factoring with oblimin rotations) and exploratory bifactor modeling using the Schmid-Leiman (SL) orthogonalization or target pattern rotation (Reise et al., 2010).

As pointed out by Reise et al. (2010), measures are rarely strictly unidimensional for broad and complex constructs. Given the need of unidimensional models due to their simplicity, "essential unidimensionality" was proposed as a weak form of the local independence assumption (Stout, 1987). Nonparametric DIMTEST can be used to test essential unidimensionality (Koğar, 2018). If a measure is sufficiently unidimensional, there would be no need to compare multiple multidimensional models.

Reise (2012) proposed two indices obtained from bifactor models to indicate degree of unidimensionality, which were the explained common variance (ECV) and percentage of uncontaminated correlations (PUC). To be specific, the ECV is the proportion of common variance among all the items attributed to the general factor, which reflects the strength of the general factor to the specific factors. The correlations among the items within each specific factor are considered to be contaminated by both common variance explained by the general factor and common variance explained by the specific factor. The number of uncontaminated correlations equals to the total number of correlations among

all the items minus the number of contaminated correlations, and the PUC is the ratio of the number of uncontaminated correlations to the total number of correlations. Larger ECV and PUC are desirable when forcing potentially multidimensional data into a unidimensional model. As recommended by Reise (2012), if test developers plan to have a unidimensional model of a relatively broader construct, they can improve PUC by increasing the number of testlets and decreasing the number of items within each testlet. In the simulation study of Reise, Scheines, Widaman, and Haviland (2013) conducted within the SEM framework, they explored the effect of misspecification of bifactor data using a unidimensional model on the structural coefficient of predicting a latent criterion from the general factor. They found that the ECV and PUC of the generated data were good predictors of estimation bias of the structural coefficient. To be specific, the estimation bias decreased as ECV and PUC increased, and the effect of ECV on the estimation bias was moderated by PUC. When the PUC was high, the structural coefficients are almost unbiased even if the ECV is low. Reise et al. (2013) also pointed out that the model fit indices (i.e., CFI, SRMR, and RMSEA) performed poorly in testing unidimensionality because the misspecified unidimensional models had acceptable model fit in most of the cases, and that these model fit indices cannot serve as predictors of the bias of the structural coefficient when predicting external criterion from the general factor. Thus, they suggested that researchers should use ECV and PUC in addition to overall model fit indices to determine the degree of unidimensionality.

When determining whether it is appropriate to fit a unidimensional model to potentially multidimensional data, another important aspect to consider is the degree of distortion of the item parameters due to forcing the data onto only one dimension. Reise et

al. (2007) pointed out that the unidimensional factor might be pulled toward the subset of items with strong local dependence. In practice, the item discrimination parameters for a unidimensional model are often compared with the corresponding item discrimination parameters of the general factor in a bifactor model to determine whether distortion occurs (Reise et al., 2007). DeMars (2006) found that the item discrimination parameters would be negatively biased if generated bifactor data was fitted with a unidimensional model based on a simulation study. Also, she indicated that fitting a complex model (i.e., bifactor model) to a simple data structure (i.e., unidimensional data) would not produce any bias but it would slightly increase root mean square error (RMSE) in the estimates of item parameters. Given there was no estimation bias and the decrease in estimation efficiency was very small when specifying a bifactor model to the unidimensional data, DeMars (2006) suggested that a bifactor model is preferred when suspecting multidimensionality, and she also proposed that researchers could specify specific factors for some of the testlets rather than all of them to improve the estimation efficiency.

With respect to person parameter (i.e., primary trait reflected by the general factor in the bifactor model) estimation, DeMars's study (2006) indicated that person parameter estimates obtained from different models (i.e., unidimensional model, testlet-effects model and bifactor model) were closely correlated with each other, whereas Min and He's study (2014) indicated that the correlation of the primary trait estimates between the unidimensional model and the bifactor model was only .772. The reliability of tests in estimating examinees' primary traits are also of interest to researchers. Borrowing the concepts from CTT, the test reliability can be calculated using the correlation between the estimated latent ability scores and their true values in a simulation study. The correlation

is conducted within each replication to represent the test reliability, and then the obtained test reliability is averaged over the replications. In DeMars's study (2006), it was found that within the same data generation condition, reliabilities were similar across different analysis models (i.e., unidimensional model, testlet-effects model, and bifactor model). Her results also indicated that tests with items generated using a unidimensional model had higher reliabilities than those generated testlet-based tests because the generated testlets brought in additional error when estimating the primary traits. In real data, the test reliability can be estimated as 1-(s_e^2/s_T^2), where s_e is the average standard error of the latent ability estimate across examinees, and s_T^2 is the estimated total variance of the population which equals to the sum of variance of latent ability scores obtained from EAP and error variance. DeMar's study (2006) indicated that reliability was overestimated when applying unidimensional models to bifactor data.

The testlet-effects model can be considered as a special case of the bifactor model, in which constraints can be placed on the relationship between the general factor loadings and the specific factor loadings or specific factor variances (Min & He, 2014). Testlet-effects models, in which a proportional constraint is placed on the relationship between the general factor loadings and the specific factor loadings for each testlet are commonly applied for model comparisons with unidimensional models and bifactor models (e.g., DeMars, 2006; Min & He, 2014), and this type of the testlet-effects model is equivalent with the second-order model (Rijimen, 2010).

The second-order model is also a restricted version of the bifactor model, in which a second-order factor is specified to explain the relationship among the first-order factors. As pointed out by Chen et al. (2006), second-order factor models are desirable when there

are substantial correlations among the first-order factors and it is hypothesized that these correlations can be accounted for by a higher-order factor. In second-order models, the second-order factor influences each item via the first-order factor. If the direct effects of the second-order factor on the items are modeled, then this second-order model would be equivalent with the bifactor model (Chen et al., 2006). Chen et al. (2006) illustrated the similar interpretations between the second-order factor model and the bifactor model from the following aspects. First, the second-order factor of the second-order factor model corresponds to the general factor of the bifactor model; second, the disturbances of the first-order factors in the second-order factor model correspond to the specific factors in the bifactor model; third, orthogonality among the disturbances of the first-order factors and the second-order factor in the second-order factor model corresponds to the orthogonality among the specific factors and the general factor in the bifactor model, and fourth, in the special case of nonexistence of a subdomain (i.e., the loadings on this subdomain are very small in a bifactor model), the disturbance of the corresponding first-order factor would be around 0 in the second-order factor model and the corresponding specific factor should not be specified in the bifactor model. As suggested by Chen et al. (2006), in addition to having less restrictions, bifactor models are preferred over second-order factor models because the interpretation and utility of the specific factors in bifactor models are more direct than using the disturbances of the first-order factors in second-order factor models. For example, employing the bifactor model, one can estimate the unique influence of each subdomain on the indicators and the unique contribution of each subdomain to an external variable (e.g., Chen et al., 2006; Gustafsson & Balke, 1993).

Another alternative to the bifactor model is the correlated-factors model which is a non-hierarchical multidimensional model. In the correlated-factors model, each item loads on only one of the dimensions and the dimensions might be correlated with each other. The correlated-factors model can be considered as nested within the bifactor model, in which only the specific factors are modeled and the covariance among items of different testlets are explained via the correlations among dimensions. As mentioned earlier, in the study of Reise et al. (2007), most of the items with fairly strong loadings in the correlated-factor models actually discriminated the general factor better than their respective specific factors in a bifactor model, which means that in the correlated-factors model the effect of each dimension on the items might be confounded with the impact of an unmodeled general factor. As suggested by Reise et al. (2007), when dimensions in the correlated-factors model are uncorrelated with each other, one could simply fit several separate unidimensional models to the data; when these dimensions are highly correlated with each other, it might indicate the existence of a general factor. To be specific, if the correlations among the dimensions are moderate (i.e.,.1 to .4), it is likely that the general factor loadings are small whereas the specific factor loadings are large, and thus a correlated-factors model is recommended; if the correlations among dimensions are higher than .40, a bifactor model would be preferred (Reise et al., 2007).

In summary, the bifactor model plays an important role in exploring dimensionality issues because it is a more general model for its competitive models. When suspecting multidimensionality, researchers are advised to estimate the degree of unidimensionality using a bifactor model. Also, bifactor models can be applied to determine which form of multidimensional models is preferred to represent the construct.

Introduction to Differential Item Functioning (DIF) and Latent Mean Comparisons

This section focuses on multiple-group models for ordered-categorical data. I first give a brief introduction to the issues of measurement invariance and DIF. Then I discuss methods to detect DIF and compare latent means within both IRT and CFA frameworks.

Brief Introduction to Measurement Invariance and DIF

Measurement invariance is considered to hold if a test measures the construct of interest the same way across different conditions (e.g., different groups of people, different occasions, different time points, etc.) that are irrelevant to the construct to be measured (Millsap, 2011). The necessary and sufficient condition for measurement invariance is (Mellenbergh, 1989; Meredith & Millsap, 1992; Millsap, 2007):

$$P(X \mid W, G) = P(X \mid W) \tag{18}$$

where X denotes observed scores, W denotes underlying latent variables, G denotes the group membership, $P(X \mid W, G)$ is the conditional distribution of observed scores on the latent variables and group membership, and $P(X \mid W)$ is the conditional distribution of observed scores on the latent variables. As shown in Equation 18, measurement invariance means that the conditional probability of a given set of observed scores given the same level of the underlying latent variables is independent of the group membership (e.g., different groups of people, different occasions, different time points, etc.).

Measurement invariance can be tested within both CFA and IRT frameworks (e.g., Flowers, Raju, & Oshima, 2002; Kim & Yoon, 2011; Meade & Lautenschlager, 2004; Reise, Widaman, & Pugh, 1993). Under CFA framework, measurement invariance is regarded as factorial invariance and it is evaluated using a series of nested models. Under

IRT framework, measurement invariance is considered to be the invariance of item response functions (IRFs) determined by the item parameters, and it is assessed by detecting differential item functioning (DIF) at the item parameter level or the IRF level (Meredith & Teresi, 2006; Reise et al., 1993).

An item is considered to show DIF if the expected score (or probability of passing this item) of an examinee given his or her latent ability score(s) is dependent on his or her group membership (Flowers et al., 2002). Different from measurement invariance that requires no differences exist in item parameters or IRFs, DIF can be viewed as the differences in item parameters or IRFs, so it is a matter of degree rather than yes or no (Borsboom, 2006). Borsboom (2006) suggested that the influence of the bias (i.e., failure of measurement invariance) of a test on making correct statistical conclusions depends on the research scenarios. For example, measurement invariance is crucial when making inferences from the differences in observed scores across groups to their latent mean differences. However, if the size of bias is much smaller than the target effects (i.e., observed score differences), there would be not a concern with the bias. Thus, as recommended by Borsboom (2006), researchers should test DIF rather than just assume measurement invariance when comparing mean differences between groups.

There are two types of DIF, uniform DIF and non-uniform DIF. Uniform DIF occurs when an item always favors one group regardless of the latent ability levels, whereas non-uniform DIF occurs when IRFs (or expected score functions) of different groups cross over at some point such that the probability of passing this item (or expected score of this item) is higher for one group for some latent ability levels and other groups at other latent ability levels. With respect to item parameters, discrimination parameters remain the same

(i.e., only difficulty or threshold parameters vary) across groups for the case of uniform DIF; discrimination parameters differ across groups for the case of non-uniform DIF (Flowers et al., 2002; Teresi, 2006). In comparison with uniform DIF, the power to detect non-uniform DIF is lower using some of the DIF detection procedure, and non-uniform DIF might have no impact on observed score differences because the item with non-uniform DIF favors different groups depending on the latent ability levels (Tay, Meade & Cao, 2015). Although the influence of non-uniform DIF on observed score differences between groups might not be obvious, it is still a concern for individuals' observed score interpretation across different groups.

In addition to DIF, differential functioning also occurs at test level. If an examinee's total expected score in a test given his or her latent ability scores is dependent on the group membership, it is said that this test shows differential functioning (Raju, van der Linden & Fleer, 1995). Borsboom (2006) pointed out that a test might be unbiased in the presence of DIF for multiple items because DIF of different items might be canceled out.

Methods of Detecting DIF

For multiple-group ordered-categorical data, measurement invariance can be explored under both IRT and CFA frameworks. Under the IRT framework, measurement invariance is usually tested by detecting differential functioning at item level (i.e., DIF). Before detecting DIF, the dimensionality issues (i.e., factor structure) need to be tested across groups (Tay et al., 2014). Also, model selection might be needed if researchers are uncertain which model best fit the data. Once the factor structure and the model form are supported as being the same across groups, the next step is to detect DIF. The most commonly applied method of detecting DIF derived from IRT models is likelihood ratio

test (the LR test; e.g., Kim & Yoon, 2011; Meade & Lautenschlager, 2004), which is also most closely related to CFA procedure involving comparisons among series of nested models. The LR test compares the likelihood values between the baseline model (M_0) and a more (or less) constrained model (M_1). Supposing that M_1 is a more constrained model, a G^2statistic can be obtained from the following formulas:

$$LR = \frac{L_{M_1}}{L_{M_0}} \tag{19}$$

$$G^2 = -2\ln(LR) = -2\ln(L_{M_1}) + 2\ln(L_{M_0}) \tag{20}$$

where LR denotes likelihood ratio, L_{M_1} is the likelihood function of M_1 and L_{M_0} is the likelihood function of M_0, which can be obtained using multiple-group MML estimation. G^2 statistic has an approximate χ^2 distribution with degree of freedom equal to the difference of parameters between M_0 and M_1 under the null hypothesis such that a significance test can be conducted for model comparisons.

When detecting DIF using LR tests, there are two approaches in selecting a baseline model: forward and backward procedures (Kim & Yoon, 2011). In the forward procedure, the baseline model is the least constrained model in which equality constraints are only placed on the anchor item(s) necessary for model identification. Then a more constrained model in which equality constraints are added on a studied item of the focal group is compared with the baseline model. In the backward procedure, the baseline model is the most constrained model in which equality constraints are placed on all items. Then a less constrained model in which equality constraints are relaxed for a studied item of the focal group is compared with the baseline model. The drawback of the forward procedure is that the selection of anchor variable(s) is somewhat arbitrary, while the backward procedure has the limitation that the baseline model is very likely to be a misspecified model in which

equality constraints are placed on noninvariant items (Kim & Yoon, 2011). In practice, the applications of LR tests are quite flexible. Researchers can combine ideas of both the forward procedure and the backward procedure. For example, researchers can use the model with equality constraints on several anchor items as the baseline model and then test DIF of a subset of items of interest. If the more constrained model significantly reduces the model fit, then equality constraints can be relaxed one by one to figure out the specific parameters with non-invariance (Millsap, 2011).

Take the unidimensional IRT model as an example. The most commonly applied identification method for multiple-group IRT models is to fix the mean of the latent ability distribution of the reference group as 0 and the variance of this distribution as 1 in addition to setting equality constraints on at least one anchor item. The mean and variance of the latent ability distribution for the focal group are freely estimated such that the latent mean difference can be estimated.

In addition to focusing on the differences in item parameters, DIF can be tested by comparing the expected score functions or IRFs across different groups using the differential item and test functioning (DFIT) framework (Morales, Flowers, Gutierrez, Kleinman, & Teresi, 2006).

The issues of DIF can be evaluated under the CFA framework as well. Some researchers applied traditional multiple-group CFA models for continuous data to explore factorial invariance of a test composing of ordered-categorical items (e.g., Flowers et al., 2002; Meade & Lautenschlager, 2004). When applying traditional multiple-group CFA models, researchers need to initially test whether the variance-covariance matrices and mean vectors are the same across different groups. If the null hypothesis of no differences

is rejected, measurement invariance (i.e., invariance about the relationships between latent factors and observed measures), structural invariance (i.e., invariance of variances and covariances among latent variables), and latent mean differences can be tested using a series of nested models. Measurement invariance should be established before testing structural invariance and latent mean differences. The evaluation of measurement invariance includes tests of configural invariance (i.e., equivalent factor structures), metric invariance (i.e., equivalent factor loadings), strong invariance (i.e., equivalent loadings and intercepts), and strict invariance (i.e., equivalent loadings, intercepts and unique error variances) step by step. Researchers do not need to conduct all these tests, and the extent to which invariance is required depends on researchers' needs. For example, if researchers are interested in latent mean differences, at least partial loading and intercept invariance should be achieved (Byrne, Shavelson, & Muthén, 1989). Like multiple-group IRT models, traditional multiple-group CFA models are usually identified by setting the latent means of the reference group as 0 and the corresponding variances as 1 in addition to placing equality constraints on the loading and intercept of at least one item per factor. There are three major drawbacks of applying traditional multiple-group CFA models for continuous data to explore measurement invariance for ordered-categorical items. First, traditional CFA models usually assume that the observed variables are continuous and normally distributed which is obviously violated for applications in tests composing ordered-categorical items. Second, the loadings in traditional CFA models can be considered as corresponding to the discrimination parameters of IRT models but no parameters of traditional CFA models correspond directly to the difficulty parameters of IRT models. Although the intercept parameters of CFA models are similar to item difficulty parameters of IRT models, they

still function differently. For example, there are more than one step difficulty parameters (or threshold parameters) for polytomous data in IRT models, but there is only one intercept parameter if the data are forced to fit with traditional multiple-group CFA models (Meade & Lautenschlager, 2004). Third, traditional CFA models describe linear relationships between observed variables and latent common factors, whereas IRT specifies a nonlinear function between the probability of examinees' responses and latent ability scores.

Although traditional multiple-group CFA models for continuous data have limitations in exploring measurement invariance for ordered-categorical data due to their fundamental differences from IRT models, multiple-group categorical CFA models can be appropriately applied for detecting DIF because of their equivalence with 2PL models and GRMs (Kim & Yoon, 2011).

The identification methods of the least constrained multiple-group categorical CFA models (i.e., baseline models) were introduced in detail by Millsap and Yun-Tein (2004). Both the factor structure (i.e., no cross loadings vs. existence of cross loadings) and the item type (i.e., binary item vs. polytomous item) can influence the identification strategy for the baseline model of multiple-group categorical CFA models. Take the case in which each item loads on only one factor and the number of categories for each item is larger than 2 as an example. One of the identification methods for the baseline model in this case is: (1) fix the means of factors in the reference group zero; (2) in each of the two groups, for each factor, select a reference variate and fix its loading to 1; (3) fix the variance of each latent continuum underlying each observed ordered categorical variable to 1 in the reference group; (4) fix all of the intercepts to zero in both groups; (5) constrain each respective threshold of each latent response continuum (e.g., the first threshold) to be equal

across groups; and (6) for the reference variate of each factor, fix an additional threshold (e.g., the second threshold) to be equal across groups. Once the baseline model is established, the invariance for factor loadings and threshold parameters can be examined through a series of nested models. The WLSMV estimator can be used for multiple-group categorical CFA models, and the DIFFTEST option in Mplus can be applied for chi-square difference tests using WLSMV estimator (Asparouhov, Muthén, & Muthén, 2006)). DIFFTEST is implemented based on the T_3 chi-square difference correction which is considered to perform better than previously proposed correction methods given its statistical properties are more similar to those for a χ^2 statistic. The DIFFTEST option has been available in Mplus since version 6, in which the model information of the two nested models is extracted to calculate the T_3 statistic via a series of complicated formulas. Like in the traditional multiple-group CFA models and multiple-group IRT models, the freely estimated latent factor means of the focal group in multiple-group categorical CFA models reflect the latent mean differences that might be of interest.

In addition to these multiple-group methods, multiple samples can be combined into one dataset and the DIF can be directly estimated by modeling group membership as a variable together with the person ability variable (either the total scores or latent ability scores) to predict examinees' responses using logistic regression or multiple-indicator-multiple-causes (MIMIC) models.

Another type of DIF detection procedure is based on nonparametric models in which examinees' performance of the reference group and the focal group are compared conditional on their abilities, and their abilities are obtained using observed scores rather than latent ability scores. Common nonparametric methods include the Mantel-Haenszel

χ^2 method (M-H), standardization method of Dorans and Kulick (1986), and SIBTEST (Shealy & Stout, 1993).

The Application of Latent Mean Comparisons for Ordered-categorical Data

Latent mean comparisons become popular within the CFA framework (Schmitt & Kuljanin, 2008). Measurement invariance needs to be tested before conducting latent mean comparisons. As suggested by Byrne et al. (1989), one could estimate latent mean differences across groups if partial invariance regarding factor loadings and intercepts is achieved. However, there are no consistent opinions about the extent to which partial factorial invariance is allowed. Some researchers suggested that for each factor there should be at least one more indicator in addition to the referent indicator having invariant factor loadings and intercepts across groups to make latent mean comparison meaningful (e.g., Steenkamp & Baumgartner, 1998), whereas some argued that the majority of items should have invariant loadings and intercepts to avoid the arbitrariness of latent mean comparisons (e.g., Reise et al., 1993).

When the indicators are subscale scores which can be considered as being from a multivariate normal distribution (e.g., Hong, Malik, & Lee, 2003), it is appropriate to apply a traditional multiple-group CFA model to estimate latent mean differences. However, if the indicators are ordered-categorical items rather than subscale scores, it might be more appropriate to estimate latent mean differences based on IRT models or categorical CFA models. Although some researchers applied traditional multiple-group CFA models for these ordered-categorical indicators (e.g., Flowers et al., 2002; Meade & Lautenschlager, 2004; Steenkamp & Baumgartner, 1998), it should be noted that violation of the multivariate normal distribution assumption of the indicators might result in unexpected

results in analysis. Given that ordered-categorical items are very common in psychological measures (e.g., measures using Likert-type scales), understanding methods for conducting latent mean comparisons based on IRT models or categorical CFA models is critical.

When conducting latent mean comparisons under IRT framework, LR tests are most commonly applied (e.g., Bolt, Hare, Vitale, & Newman, 2004; Jeon et al., 2011; Oishi, 2006; Woods, Cai, & Wang, 2012). In addition, Wald tests, MIMIC and hierarchical IRT models can be also conducted for estimating latent mean differences based on IRT models (e.g., Woods et al., 2012; Finch, 2005; Jong et al., 2007). In practice, observed score differences are commonly used for making inferences. In comparison with making inferences based on observed group mean differences, conducting latent mean comparisons based on IRT models offer the following two major advantages.

First, when applying IRT models, items with larger discrimination parameters contribute more to the latent ability estimation such that in comparison with the summed total scores, the latent ability scores are more closely correspond to the true scores (Oishi, 2006). In other words, the latent mean difference reflects more pure true score difference.

Second, the observed mean differences are the combination of latent mean differences and test level differential functioning. For example, in the study exploring the gender difference on a stress reaction measure (Smith & Reise, 1998), the observed gender differences in the stress reaction measure reflected both the latent mean difference in the negative affectivity factor and the gender differences in expressions of the negative affectivity. Given that the focus of the research is the mean differences in the target construct, making inferences based on observed score differences would produce confounding effects. In contrast, latent mean comparisons are conducted based on invariant

items which can avoid such confounding effects. It should be noted that the summation of items that might show DIF does not necessarily lead to biased observed differences because DIF of different items might cancel out each other. In the above example, some of the items might be easier for women to endorse whereas some might be easier for men to endorse, and thus, as suggested by Borsboom (2006), blind removal of items with severe DIF might induce more bias at the test level.

As mentioned earlier, measurement invariance needs to be tested before conducting latent mean comparisons using multiple-group CFA models. Similarly, when conducting latent mean comparisons based on multiple-group IRT models, DIF needs to be detected first for item parameters (e.g., Bolt et al., 2004; Oishi, 2006). In practice, the DIF of the item parameters are probably due to the following reasons: (1) examinees' understanding of the concepts might not be identical across groups; (2) the translation of the measure might be improper; (3) the examinees in some groups might avoid extreme responses; (4) social desirability or social norms might differ across groups; (5) some of the items might be more easier for a given group than for other groups; (6) examinees' of different groups might have different reference points when describing themselves (Chen, 2008). From a measurement standpoint, the presence of DIF might be due to group differences in the unmodeled common factor(s) or systematic errors (Meredith & Teresi, 2006).

Additionally, as when conducting latent mean comparisons within the multiple-group CFA framework, latent mean differences can be estimated based on partial invariant item parameters of the multiple-group IRT model (e.g., Bolt et al., 2004; Oishi, 2006). For example, in the study of Oishi (2006), the mean difference in the Satisfaction with Life Scale (SWLS) was estimated between American and Chinese samples. In their study based

on a multiple-group unidimensional model, only 1 item was shown to be invariant across groups after conducting a series of LR tests and the latent mean comparison was conducted with only this invariant item constrained to be equal across groups. Oishi (2006) suggested that more invariant items are needed to obtain a truly unbiased estimate of the latent mean difference although one could get a solution of the latent mean difference using only one anchor item of the latent factor.

Factors Influencing DIF Detection and Latent Mean Comparisons

In this section, I illustrate the factors that influence DIF detection and latent mean comparisons based on the findings of a variety of simulation studies.

There are two types of factors that influence the detection of DIF, those regarding the data generation process (e.g., sample size, effect size of DIF, item parameters, person parameters, etc.), and those regarding the data analysis process (e.g., DIF detection methods, data analysis model). In addition to the main effect of each factor on the DIF detection procedure, the joint effect of several factors on how well DIF can be detected are also of interest to researchers. The outcomes that are focused on are usually the Type I error rate and power in detecting DIF.

Factors Influencing Type I Error Rates for DIF Detection

Acceptable Type I error rates are the prerequisite for making inferences regarding empirical powers. As shown in Table 1A, the data generation factors influencing Type I error rates are the sample size, the data generation model, the percentage of DIF, the pattern of DIF across items, and the distributions of person parameters; the data analysis factors influencing Type I error rates are the model selected for DIF detection (e.g., whether or not misspecified) and the methods used for DIF detection (e.g., nonparametric vs. parametric

methods, forward approach vs. backward approach, and whether or not using Bonferroni corrections).

In the study of Cohen, Kim & Wollack (1996), the datasets were generated using either the 2PL model or the 3PL model, and no DIF was generated for the items. When analyzing the data, the datasets generated using the 2PL model were correctly specified, and the datasets generated using the 3PL model were either correctly specified or misspecified by fixing the pseudo-guessing parameter to the average value of the pseudo-guessing parameters of all the items. The DIF was tested for each item sequentially using other items as anchor items, and the proportion of significant LR test results across all the replications across all the items for each research condition reflected the Type I error rate for this condition. Their results indicated that the Type I error rates were close to the nominal alpha level for the 2PL model conditions, and the Type I error rates were a little higher for the 3PL model conditions, especially when the nominal alpha level was at .0005 to .005. Also, it indicated that sample size did not influence Type I error rates obviously in their research scenario.

Bolt's study (2002) showed that slight misspecification of the model would lead to large inflation of Type I error rates when applying LR tests to detect DIF for the items analyzed using graded response models (GRMs), and the Type I error inflation was especially severe when the sample size was large (i.e., 1000 in each group). It also indicated that there were much less Type I error rate inflation due to model misspecification if using DFIT for DIF detection, and that the Poly-SIBTEST (a nonparametric estimation method; Chang, Mazzeo, & Roussos, 1996) seemed to be unaffected by the generating models in terms of Type I error rates. Thus, as suggested by Bolt (2002), when the sample size is

large, researchers should be cautious about use of LR tests to detect DIF for items of the GRM when they are uncertain if the GRM properly represents the construct.

Type I error rates are also influenced by the distributions of person parameters. In the study of Ankenmann, Witt, & Dunbar (1999), both LR tests and the Mantel procedure (a nonparametric estimation method) showed good control over Type I error rates when the distributions of ability parameters were identical across groups. However, when the latent mean difference between the reference group and the focal group was nonzero, LR tests still maintained acceptable control over Type I error rates whereas the Mantel procedure lacked control over Type I error rates.

When detecting DIF using LR tests based on multiple-group IRT models or traditional multiple-group CFA models, the strategy in setting up baseline models (forward approach vs. backward approach) have a great influence on Type I error rates (e.g., Stark, Chernyshenko, & Drasgow, 2006). As found by Stark et al. (2006), when using the constrained-baseline model (backward approach), both LR tests under IRT models and chi-square difference tests under CFA models showed substantial Type I error inflation unless no DIF existed in the fully constrained model, and the Type I error inflation could be reduced by applying a Bonferroni-corrected critical *p* value. As suggested by Wang and Yeh (2003), when conducting LR tests using all other items as anchor (constrained-baseline model), Type I error inflation occurred when the percentage of items with DIF reached 12% under the 3PL model and 20% under the 2PL model and the GRM for the conditions in which all the items with DIF favored one group (one-side conditions), whereas for the conditions in which some of the items favored the reference group while some favored the focal group (both-side conditions), the performance of the constrained baseline model in

controlling over Type I error rates was determined by average signed area (i.e., the average difference between IRFs of each item). The larger the average signed area was, the more severe Type I error inflation produced by using constrained baseline model was for both-side conditions.

Factors Influencing Power for DIF Detection

Once acceptable Type I error rates are achieved, researchers can appropriately interpret power in simulation studies. As shown in Table 2A, the data generation factors that influence the power in detecting DIF include sample size, the ratio of sample size in the reference group to the focal group, the effect size and pattern of DIF, the magnitude of item parameters, the item type (binary vs. polytomous), and the distributions of person parameters; the data analysis factors that influence the power in detecting DIF include the model selected for DIF detection and the methods to detect DIF (e.g., nonparametric vs. parametric methods, IRT-based methods vs. CFA-based methods, forward approach vs. backward approach, the number of anchor items, and whether or not using Bonferroni corrections).

As expected, many studies based on different DIF detection methods have shown that the power in detecting DIF increases as the sample size increases (e.g., Ankenmann et al., 1999; Kim & Cohen, 1992; Raju, Drasgow, & Slinde, 1993). Generally, large sample size is required to obtain accurate item parameters for IRT models. To recover item parameters with little bias for a 2PL model, the sample size of 500 is usually required for a test with less than 40 items (e.g., Reise & Yu, 1990; Stone, 1992) although this requirement was not satisfied in some of the simulation and real data research. For a 3PL model, more examinees are needed to obtain accurate estimation of item parameters. The

study of Ankenmann et al. (1999) suggested that LR tests lacked power in detecting DIF when the sample size was as small as 500 in each group.

In addition to total sample size, the ratio of sample size between the reference group and the focal group also influences the power in detecting DIF. The results of Sweeney's study (1996) indicated that for a given total sample size, the power to detect DIF was higher for equal sample size conditions than the conditions with much fewer examinees in the focal group.

It can be also expected that effect size of the DIF would influence the power to detect it. Previous studies have consistently indicated that the items with larger effect size of DIF were more easily detected as showing DIF (e.g. Narayanan & Swaminathan, 1996; Sweeney, 1996). As suggested by Borsboom (2006), the ratio of effect size of DIF to the latent mean difference of the person ability scores is crucial in judging whether such DIF should be paid attention to.

The power in detecting DIF is not only affected by the effect size of the DIF but also affected by the pattern of the DIF. In the study of Ankenmann et al. (1999), noninvariant threshold parameters for GRMs were simulated. In the constant DIF pattern, a value was added to each of the threshold parameter of the noninvariant item in the reference group to obtain the threshold parameters of this item in the focal group, which corresponds to the practical condition where an item is more difficult for the focal group than the reference group. In the balanced DIF pattern, to obtain the threshold parameters of the noninvariant item in the focal group, a value was added to the lowest threshold parameter of the noninvariant item in the reference group while the same value was subtracted from the highest threshold parameter of this noninvariant item in the reference

group, which corresponds to the practical condition where examinees in the focal group tend to avoid extreme responses. Ankenmann et al. (1999) found that the Mantel procedure (a nonparametric estimation method) showed greater power than LR tests for the constant DIF pattern conditions when the person ability distributions were identical across groups. However, for the balanced DIF pattern conditions, LR tests showed much higher power than Mantel procedure.

As suggested by Bolt's study (2002), in comparison with the nonparametric DIF detection method (i.e., Poly-SIBTEST), parametric DIF detection methods such as LR tests and DFIT showed greater power in detecting DIF for the items of GRMs when the model was correctly specified, and this advantage was especially obvious for the small sample size conditions (i.e., 300 for each group). Also, when the sample size was small (i.e., 300 in each group), these parametric methods also showed acceptable Type I error rates even under conditions of slight model misspecification. Thus, it was concluded that the parametric methods are preferable for DIF detection when the sample size is small.

Sweeney (1996) found that the magnitude of item parameters influenced the power to detect DIF for them, which means that for the same amount of DIF, it might be detected in one item but not in another item. Sweeney (1996) explored the joint effects of the ratio of reference group sample size to the focal group sample size, effect size of the DIF, the magnitude of item parameters and the person ability distributions on the power to detect DIF using LR tests and concluded that the power to detect DIF depended on the following two joint factors: (1) the differences between the IRFs for the reference group and the IRFs for the focal group; (2) the number of focal group examinees located on the latent ability continuum where the IRFs differ across groups. For example, if an item is too easy or too

difficult relative to examinees' abilities, the difference of IRFs between the reference group and the focal group would be almost zero over most of the ability range such that the DIF in difficulty parameters would be difficult to be detected.

In comparison with the traditional multiple-group CFA models, although IRT based methods do not have the problems of violations of normality and continuity, they require larger sample sizes than CFA models to achieve a given degree of accuracy in locating the items with DIF. Based on the results of Stark et al. (2006), it was recommended to apply traditional multiple-group CFA models for detecting items with DIF when the sample size is small and the items are polytomous. Also, Stark et al. (2006) pointed out that the free-baseline models (forward approach) performed better than the constrained-baseline models (backward approach) for both LR tests under IRT models and chi-squire difference tests under CFA models. In addition, although Bonferroni corrections were helpful in reducing Type I error inflations for the backward approach, they were not recommended for small samples because they would reduce power as well due to their strict criterion in obtaining significant results.

Although the constrained-baseline model leads to substantial Type I error inflation, using some of the items as anchor usually yields good control over Type I error rates (Wang & Yeh, 2003). In Wang and Yeh's study (2003), they simulated 25 items. They pointed out that using 1 anchor item could appropriately control over the Type I error rate while showing acceptable power of detecting DIF, and that using 4 or 10 anchor items would lead to higher power. Thus, researchers can select a baseline model that is more constrained than free-baseline model but less constrained than the constrained baseline model. As suggested by Wang and Yeh (2003), anchor items can be selected based on

related theories (or experts' opinions) and preliminary analyses. The LR tests using all other items as anchor is one of the commonly applied methods to locate anchor items.

Simulation Studies for Multiple-group Bifactor Models

To the best of my knowledge, only three simulation studies have focused on DIF detection under bifactor IRT models. Two of these three studies (e.g., Cai, Yang, & Hansen, 2011; Jeon, Rijmen, & Rabe-Hesketh, 2013) were conducted using extended methods of full-information marginal maximum likelihood estimator with dimension reduction technique (Gibbons & Hedeker, 1992), and the other one (Fukuhara & Kamata, 2011) was conducted using a fully Bayesian estimation method.

Jeon et al. (2013) allowed the orthogonality assumption to be violated in the focal group, and they found that ignoring between-group differences in the relationship among latent variables resulted in substantial bias in DIF estimates. Cai et al. (2011) worked on an extended multiple-group, bifactor IRT model in which the model can be very flexible. They conducted two simulation studies. In the first study, the examinees in Group 2 did not take items related to one of the specific factors in Group 1, which corresponded to the realistic condition where existing group specific subdomain(s). In the second study, the generated data for both groups consisted of multiple types of items (i.e., multiple-choice items, constructed response items, complex multiple-choice items), which corresponded to realistic educational tests. Cai el al.'s simulation study (2011) was conducted to illustrate the efficiency of the proposed estimation method. In their study, the data analysis model was consistent with data generation model, and they were interested in the recovery of parameters including latent mean differences.

Fukuhara and Kamata (2011) generated testlet-based data with DIF in difficulty parameters and analyzed data using both a bifactor IRT model and a unidimensional IRT model. Their results indicated that the bifactor model could produce better DIF detection and more accurate estimates for DIF magnitude in comparison with the unidimensional model which ignored the local dependency resulting from the testlets.

Simulation Studies for Latent Mean Comparisons

Although person ability distributions were usually generated to be different in previous simulation studies exploring the DIF issues, in most of the studies, the latent mean difference was treated as a factor that might have an impact on DIF detection procedure rather than the outcome (e.g., Stark et al., 2006).

In some of the simulation studies about multiple-group IRT models, the estimation of accuracy for the latent mean difference was focused on (e.g., Kim & Cohen, 1998; Woods et al., 2012). In these studies, latent mean differences were estimated based on anchor items generated to be invariant, and no estimation bias for the latent mean differences was found. One possible explanation was that there would be no bias as long as there was no misspecification.

Similar results were also found in some CFA-based simulation studies. Yang (2008) systematically explored the influence of partial loading invariance and partial intercept invariance as well as some other important factors on latent mean comparisons under CFA framework. In this study, all models were correctly specified, and the results indicated that the power of detecting latent mean differences was higher for the complete invariant model than the model with noninvariant components and, further, that the power was not influenced by the degree of noninvariance. For all research conditions, there was no

obvious estimation bias, which might be due to the same reason that all the models were correctly specified.

Hancock, Lawrence, and Nevitt (2000) explored the impact of misspecification of the model by setting equality constraints to noninvariant loadings on latent mean comparisons based on multiple-group CFA models. They also varied sample size, between-group sample size ratio, the pattern of sample size, and population generalized variance. They found that the power to detect the latent mean difference was lower when the model was misspecified by ignoring the noninvariance in comparison with the corresponding correctly specified model. It also indicated that in most of the cases, larger disparity between sample sizes for the groups was associated with decreased power to detect the latent mean difference. Additionally, in comparison with the conditions in which the group with larger sample size was associated with smaller population generalized variance, the power to detect the latent mean difference was lower for the conditions in which the group with the larger sample size was associated with larger population generalized variance.

Beuckelaer and Swinnen (2018) generated a two-group single-factor CFA model with 3 or 4 indicators, with the second indicator having noninvariant loading or intercept in some research conditions, and then the latent mean difference was estimated based on the traditional multiple-group CFA model assuming strong invariance was achieved. They also manipulated the type of distribution of the items (normal distribution, discrete 5-point scales with either a unimodal left-skewed distribution or a symmetric bimodal distribution), sample size in each of the group, effect size of latent mean difference, and noninvariance of factor loadings or indicator intercepts. Their results indicated that ignoring noninvariance may have a very strong influence on the percentage of correct statistical

conclusions of the latent mean difference tests. The probability of drawing correct statistical conclusions of the latent mean difference tests was strongly reduced due to ignoring the difference in the noninvariant indicator intercept of about one-tenth or even smaller of the total length of the scale. Also, ignoring the difference of 0.2 in the factor loading also reduced the percentage of correct conclusions of the latent mean difference tests. In their study, sample size and the distribution of the indicators did not influence the percentage of correct conclusions regarding the latent mean difference tests.

Jones and Gallo (2002) detected DIF and estimated the latent mean difference of Mini-Mental State Examination responses across different groups (i.e., high-education group vs. low-education group; male vs. female) using MIMIC for dichotomous items, and they also examined the effect of ignoring DIF on latent mean difference estimates by purposely fixing the direct effects of group membership on the response variates to zero. The MIMIC for dichotomous items approximates to 2PL IRT models with discrimination parameters assumed to be equal across groups. The direct effects of group membership on the response variates reflect the differences in threshold parameters (or difficulty parameters) across groups. In their study, there were 31 items in total, 10 of them showed DIF between high-and low-education groups while 16 of them showed DIF between male and female. They found that ignoring DIF resulted in 1.6% overestimation of the latent mean difference between high- and low-education group, and 95% overestimation of the latent mean difference between male and female.

The Purpose of the Current Study

Bifactor models have gained increasing popularity in recent years because they often serve as the most appropriate representations for relatively broader psychological

constructs containing multiple narrower aspects and testlet-based cognitive tests including items based on common stimulus. When researchers are interested in comparing such complex psychological constructs or testlet-based cognitive tests among multiple populations, they need to rely on multiple-group bifactor models. From a methodological standpoint, however, only a few studies have focused on multiple-group bifactor models, and to the best of my knowledge, all of these studies involved the factors that influenced DIF detection of the item parameters (e.g., Cai et al., 2011; Fukuhara & Kamata, 2011; Jeon et al., 2011). Given that researchers are usually very interested in the primary trait represented by the general factor when applying bifactor models and they might be also interested in comparing latent means of the primary trait among multiple populations, the main purpose of the current study is to systematically explore the factors that influence the latent mean comparisons of the general factor for bifactor ordered-categorical data.

Bifactor models have been widely applied for the purpose of exploring dimensionality issues in single-group analysis. With bifactor models, the degree of unidimensionality can be estimated such that researchers can have more information regarding the consequence of violation of local independence assumption. Although multiple methods are employed in determining unidimenisionality in practice, the criterion is somewhat arbitrary. Also, unidimensional models might be preferred to multidimensional models in some cases for the purpose of theoretical simplicity. Thus, the first specific aim of the current study is to explore the impact of fitting the bifactor ordered-categorical data using unidimensional models on the latent mean comparison for the general factor.

In real data analysis, most of multiple-group comparisons using bifactor models are based on traditional CFA models. On one hand, for the items with only a few categories, the applications of traditional CFA models might not be appropriate because of violation of normality and continuity assumptions. On the other hand, as pointed out by Stark et al. (2006), traditional CFA models might be preferred to IRT models for polytomous data when the sample size is small because IRT models in which more parameters need to be estimated require larger sample size than the corresponding traditional CFA models to achieve certain degree of estimation accuracy. Thus, the second aim of the current study is to explore the impact of treating ordered-categorical data as continuous data under varied conditions in bifactor model framework.

Although complete measurement invariance is ideal for latent mean comparisons, it is hard to achieve in application. In most cases, latent mean comparisons are conducted under partial measurement invariance obtained from stepwise selection of noninvariant parameters. The post hoc adjustments on multiple-group models have been criticized by many researchers (e.g., Marsh et al., 2018), and previous simulation study results also suggested that the noninvariant parameters can never be perfectly recovered under traditional multiple-group analysis based on such post hoc manner unless the effect size of the DIF or the sample size was very large. Although an alternative method named alignment method (Asparouhov & Muthén, 2013) has been proposed recently to avoid the problems of the commonly applied post hoc selection of noninvariant parameters, it cannot be implemented for bifactor model cases because this new method only applies to the model in which each indicator loads on only one factor. Thus, in order to better interpret the latent mean difference of the general factor of a bifactor model, it is necessary to

consider the possibility of failing to account for DIF. Accordingly, the third specific aim of this study is to explore the impact of failing to account for DIF in different parameters on the latent mean comparison of the general factor for the generated multiple-group, bifactor, ordered-categorical data.

Chapter 2: Methods

Overview

Monte Carlo simulation methods were applied in the current study to explore the robustness of latent mean comparisons for the general factor of the generated bifactor, ordered-categorical data to varied research conditions. Although the IRT framework provides complete features in terms of person parameters and item parameters when latent ability is measured by ordered-categorical variables, the commonly applied IRT models (i.e., 2PL model or GRM) are equivalent to categorical CFA models and they can be estimated with categorical variable methodology in the CFA framework. In the current research scenario, the focal variable was the latent mean difference rather than each person's parameter estimate and no pseudo-guessing parameter was assumed, so theoretically speaking, both IRT models and categorical CFA models can be used to fit the data equivalently. Given that no simulation studies have focused on multiple-group, bifactor, categorical CFA, and for some researchers, SEM software (e.g., Mplus) might be relatively more easily accessible, it was chosen as the estimation method in the current study.

The representative model is a bifactor categorical CFA model, which is equivalent with a bifactor GRM within the IRT framework. A set of specific research conditions, including varying population characteristics, sample characteristics, item characteristics and data analysis strategies, were designed to address the research questions. For each research condition, 1000 entire response datasets were generated using R based on multiple-group bifactor GRMs. For some conditions, the data analysis model was different from the data generation model to study effects of model misspecification in terms of fitting

bifactor data using a unidimensional model, treating ordered-categorical data as continuous data, and setting equality constraints to the noninvariant item parameters. Mplus 7 was used for data analysis, which was implemented with *MplusAutomation* package in R. Using *MplusAutomation* package, the results obtained from fitting the data analysis model to each sample included overall model fit and parameter estimates. The results from all converged replications within a condition were summarized descriptively, including estimation bias ($E(\hat{\theta}) - \theta$), relative estimation bias ($(E(\hat{\theta}) - \theta) / \theta$), power (or Type I error rate), estimated variance, mean of the comparative fit index (CFI), mean of the standardized root mean square residual (SRMR), mean of the weighted root mean square residual (WRMR), and mean of the root mean square error of approximation (RMSEA). Details regarding the representative model, research conditions, the procedures for data generation and analysis, and the statistical outcomes of interest are illustrated in this section.

Representative Model

The representative model was a bifactor categorical confirmatory factor analysis model involving both mean and covariance structures, as shown in Figure 1, which is equivalent with the bifactor GRM within the IRT framework. For all conditions in this study, two populations were specified and the model was configurally invariant across these populations. As shown in Figure 1, a general factor was hypothesized to explain common variance among all 12 items, and each item loaded on one of the three specific factors. The general factor and all specific factors were orthogonal with each other.

For both populations, the variance of the general factor was set to 1 when generating the data, such that the generated latent mean difference can be considered as a standardized effect size. To obtain standardized interpretations of other parameters, the variances for all

specific factors were set to 1, all loadings were specified as completely standardized factor loadings, and the unit variance of each continuous latent variate underlying the corresponding ordered-categorical item was obtained through setting appropriate residual variance. To be specific, the residual of each continuous latent variate was set to one minus the sum of its variance explained by the general factor (e.g., for item 1, it is $\lambda^{*2}_{GEN1,1}$) and its variance explained by the corresponding specific factor (e.g., for item 1, it is $\lambda^{*2}_{GR1,1}$).

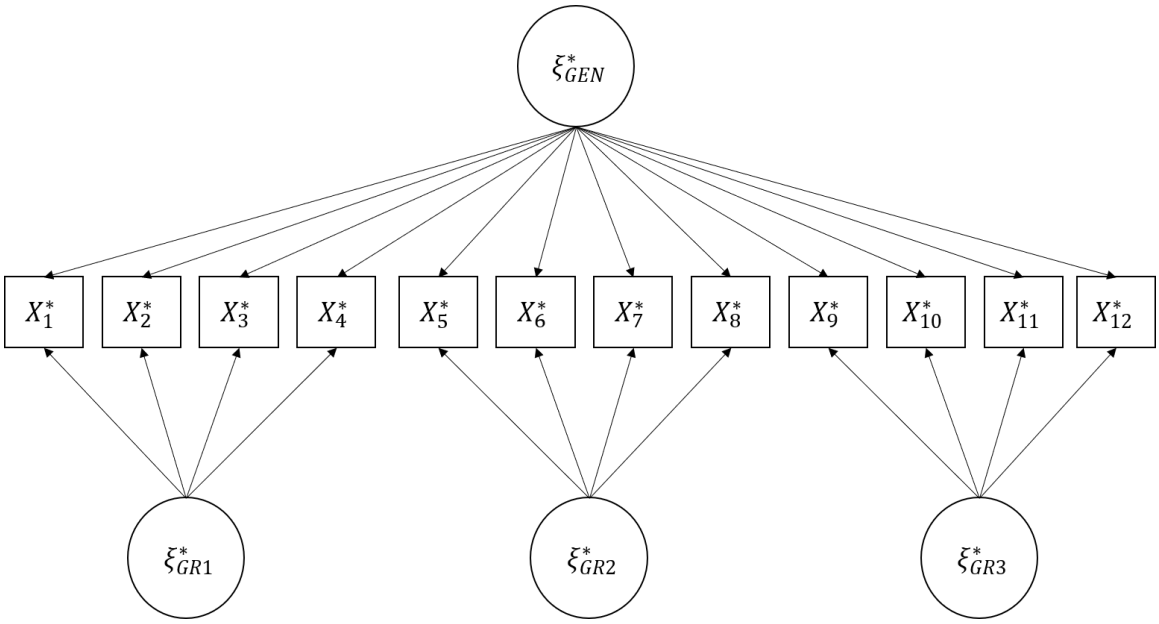

Figure 1 The Representative Model

In analysis, factors were scaled by setting the means of all latent factors to 0 and the variances of all latent factors to 1 in the reference group, and setting equality constraints for at least one factor loading for each latent factor. For bifactor models, to identify the variance-covariance matrix and mean vector among the continuous latent variates underlying the polytomous ordered-categorical items ($X^*_1 - X^*_{12}$) it was required to set the variances of the continuous latent variates to 1 in the reference group, and to constrain at

least two of the threshold parameters for each measured variable to be invariant across groups (Millsap & Yun-Tein, 2004); for binary data, in addition to setting the equality constraint on the only threshold parameter for each measured variable, the variances of the continuous latent variates were required to be set to 1 in both groups. Given that the threshold parameters and intercept parameters are indeterminate, the intercepts for both groups were set to 0. To minimize potential confounding effects on estimation bias of the latent mean difference in the general factor produced due to identification constraints, factor means were generated to be 0 in the reference group, variances of all latent factors and continuous latent variates were generated to be 1 as mentioned above, and all intercepts were generated to be 0 in both groups. For the threshold parameters on which equality constraints were applied for identification purposes, these might not be generated to be invariant across groups in some research conditions. For these conditions, it was not possible to specify the analysis models to be fully consistent with the corresponding generation models due to the needs of model identifications.

Although in realistic situations, anchor items may be chosen arbitrarily, in the current simulation study, it was assumed that the chosen anchor items were generated to have invariant item parameters across groups such that the estimation bias would be unrelated with the choice of anchor items. In the current study, the first three measured variables for each specific factor served as the anchor items and they were generated to have invariant general factor loadings, specific factor loadings, and threshold parameters across groups. When analyzing the data, all the item parameters for these anchor items were constrained to be equal across groups. The last measurement variable for each specific factor may have noninvariance in the general factor loading, specific factor loading, or

threshold parameters. When analyzing the data, the factor loadings and threshold parameters for these items were either freely estimated or constrained to be equal across groups depending on the identification requirement and research conditions.

The generated datasets were all bifactor, ordered-categorical data with either 2, 3, or 5 categories per item, and the number of categories for each item remained the same across all items and groups for each condition. The values of the parameters fell in a similar range as those from previous simulation studies and applied research involving multiple-group, bi-factor IRT models (e.g., Berkeljon, 2012; Cai et al., 2011; Fukuhara & Kamata, 2011). For all conditions, the threshold parameters (τ_j) of the reference group were set to 0 for all the items in the binary data case, they were set to -0.5 and 0.5 for all the items in the 3 categories per item case, and they were set to -0.9, -0.3, 0.3, and 0.9 for all the items in the 5 categories per item case. The general factor loadings of the reference group were set to 0.7 for all items in all conditions. The specific factor loadings of the reference group were generated to be identical across all items, but their magnitudes varied across different research conditions. In the focal group, the anchor items were generated to have the same general factor loadings, specific factor loadings, and threshold parameters as the reference group. For items with noninvariant factor loadings, the sizes of the DIF were -0.05, -0.10, and -0.15. For items with noninvariant threshold parameters, a constant value (i.e., 0.05, 0.10 or 0.15) was added to each of the threshold parameters, such that the noninvariant item consistently favored one group over the other. The latent mean difference of the general factor varied across research conditions. The latent mean differences of the specific factors were set to -0.1, 0, and 0.1 for the three specific factors, respectively, which corresponds to a realistic situation in which the focal group members and reference group

members show strengths in different specific dimensions beyond the influence of the general dimension; these values remained the same across all the research conditions.

Research Conditions

The main factors manipulated in the current study may be distinguished by those manipulated within the data generation phase and those varied within the analysis model. For the generated multiple-group bifactor IRT models without DIF, the data generation conditions varied were sample size, the number of categories per item, effect size of latent mean difference for the general factor, and the size of specific factor loadings; in the data analysis conditions, model misspecification conditions were introduced in which the generated bifactor data were fit using a unidimensional model, and/or ordered-categorical data were treated as continuous data. For the generated multiple-group bifactor IRT models with DIF, the data generation conditions varied were sample size, the number of categories per item, effect size of latent mean difference for the general factor, the type of item parameters (the general factor loadings, the specific factor loadings, or the threshold parameters) that had DIF, and the magnitude of DIF; the data analysis conditions varied in whether or not setting equality constraints on the noninvariant item parameters. The total number of specific research conditions was 408. The specific design parameters are described below and are summarized in Tables 1 and 2.

Research Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

In the following research conditions (shown in Table 1), all items were generated to have invariant item parameters across groups, and they were constrained to be equal in analysis. The data generation conditions included sample size, the number of categories

per item, size of latent mean difference for the general factor and size of specific factor loadings. The data analysis conditions included the types of analysis models and estimation methods. The specific conditions were described below.

Table 1

Manipulated Factors for the Generated Multiple-group Bifactor Models without DIF

Factors	Design Parameters
Total sample size	600 or 1200
The number of categories per item	2, 3, or 5
Size of latent mean difference for the general factor	0, -0.1, or -0.2
Size of specific factor loadings	0.3 or 0.5
The types of analysis models	Unidimensional models or bifactor models
Estimation methods	Traditional CFA models with MLR estimator or categorical CFA models with WLSMV estimator

Note: Traditional CFA models with MLR estimator was not applied for the conditions with binary data.

Sample size. Sample size is an important factor that influences the power to detect the latent mean difference of the general factor. Also, previous research (Stark et al., 2006) suggested that for ordered-categorical data with small sample size, traditional CFA might be preferred because the sample size requirement is very high for IRT models. In the current study, total sample size was varied to include 600 or 1200 cases. These sample sizes were chosen to avoid floor effects or ceiling effects regarding power, based on research conditions examined in preliminary analyses for which the difference between them was meaningful. Sample sizes were equal between the two groups.

The number of categories per item. In addition to sample size, the number of categories for each item also determines whether it is appropriate to specify a traditional CFA model for the ordered-categorical data. As the number of categories increases (i.e., 5 categories or more), in comparison with IRT models, it might be more appropriate to treat the ordered-categorical data as continuous and apply CFA models because fewer parameters need to be estimated in a CFA model for continuous data (e.g., Stark et al., 2006). In the current study, to examine the joint impact of the number of categories and the choice of estimation methods, the number of categories was set to 3 or 5 for each item, with the number of categories being the same across all items and groups for each condition. Given the popularity of dichotomous data, binary datasets were also generated, and they were only analyzed using categorical CFA models.

Size of latent mean difference for the general factor. The latent mean difference of the general factor was 0, -0.1, or -0.20. These values were chosen based on preliminary analysis to avoid floor or ceiling effects regarding power. When generating data, the variance of the general factor was set to 1 for both groups, so the corresponding latent mean difference can be considered as the standardized effect size. An effect size of 0 is consistent with the null hypothesis of no between-population differences in latent means; rejection of the null hypothesis in this case is a Type I error.

Size of specific factor loadings. As suggested by Reise (2012), the explained common variance among all items due to the general factor (ECV) can be used to indicate the degree of unidimensionality of a given dataset. In the current study, the general factor loadings of the reference group were set to 0.7 for all conditions, which is a typical value according to previous studies (e.g., Reise, 2012), so the varied size of specific factor

loadings reflected different degrees of unidimensionality of the data. The specific factor loadings of the reference group were generated to be identical across all items, and they were set to 0.3 and 0.5, respectively. The conditions with specific loadings of 0.3 corresponded to the situation containing less multidimensional data in which the specific factors have relatively small unique contribution to examinees' performance after controlling for the general factor and there might be no need to report subscale scores. When the specific factor loadings were 0.5, it corresponded to the situation containing more multidimensional data in which examinees' strengths and weaknesses on the subscales beyond the influence of the general factor was substantial.

The type of analysis models. The ordered-categorical data were generated to have bifactor structure. In applied data analysis, multiple-group IRT models might be conducted based on the unidimensionality assumption without testing dimensionality, as there is not an absolute criterion to judge unidimensionality (Tay et al., 2014). Thus, in the current study, the generated bifactor ordered-categorical data was fitted with either bifactor models or unidimensional models to explore the impact of misspecification of the model structure on the latent mean comparison of the general factor in a bifactor model.

Estimation methods. Stark et al. (2006) suggested that polytomous data (i.e., 5- or more scale points) can be analyzed as continuous data using CFA models with ML estimator when the sample size was small (i.e., 1,000 or less), given that the number of estimated parameters is relatively small for CFA models for continuous data in comparison with IRT models. Thus, in the current study, the generated bifactor, ordered-categorical data with 3 or 5 categories and no DIF were analyzed using either categorical CFA models with the WLSMV estimator or CFA models in which ordered-categorical data were treated

as continuous. Given that ML parameter estimates with standard errors and a chi-square statistic that are robust to non-normality can be obtained via robust maximum likelihood (MLR) estimator, the MLR estimator was chosen as the estimator for CFA analysis in which ordered-categorical data were treated as continuous.

Research Conditions for the Generated Multiple-group Bifactor IRT Models with DIF

In the following research conditions (shown in Table 2), the specific factor loadings in the reference group were set as 0.5. When analyzing the data, the data were fit with bifactor models using categorical CFA models with WLSMV estimator. The data generation conditions included the sample size (300 or 600), the number of categories per item (2, 3, or 5), latent mean difference of the general factor (0, -0.1, or -0.2), the type of item parameters with DIF, and the magnitude of DIF. The data analysis model varied according to whether noninvariant item parameters were constrained to be equal. The specific conditions unique to the generated multiple-group bifactor IRT models with DIF are described below.

The type of item parameters with DIF. In the current study, the types of item parameters that might have DIF included the general factor loadings, the specific factor loadings, and the threshold parameters. For each of the research condition in which DIF was generated, only one of these three types of item parameters had DIF, such that the impact of types of item parameters having DIF on the latent mean comparison of the general factor can be examined.

The magnitude of DIF. For all research conditions, only the last measured variable of each specific factor had DIF, so only three measurement variables (X_4^*, X_8^* and X_{12}^*) had

DIF. The sizes of DIF for the factor loadings were -0.05, -0.1, or -0.15. To be specific, in the research conditions in which the general factor loadings had DIF, the general factor loadings were 0.65, 0.60, or 0.55 for these three measured variables in the focal group; in the research conditions in which the specific factor loadings had DIF, the specific factor loadings were 0.45, 0.40, or 0.35 for these three measured variables in the focal group. The size of the DIF for the threshold parameters were also 0.05, 0.1, or 0.15, which was added to each of the threshold parameters with DIF in the reference group, suggesting that these items were more difficult for the focal group. The values of the DIF were informed by those used in previous studies (e.g., Fukuhara & Kamata, 2011). In the research conditions with polytomous data (i.e., 3 or 5 categories per item) in which the threshold parameters had DIF, the threshold parameters were (-0.45, 0.55), (-0.40, 0.60), or (-0.35, 0.65) for these three measured variables in the focal group when there were 3 categories per item, and they were (-0.85, -0.25, 0.35, 0.95), (-0.80, -0.20, 0.40, 1.00), or (-0.75, -0.15, 0.45, 1.05) when there were 5 categories per item. For binary data conditions, the threshold parameters were 0.05, 0.10, or 0.15 for these three measured variables in the focal group.

Equality constraints on noninvariant parameters. Given that perfect recovery of DIF is hard to achieve in practice, in the current study, the parameters that were generated to have DIF were either freely estimated (i.e., correctly specified) or constrained to be equal (i.e., misspecified) in the analysis. When testing measurement invariance using multiple-group categorical CFA models, loading invariance and threshold invariance are usually tested sequentially (Millsap & Yun-Tein, 2004). In practice, for the items with known DIF in factor loadings, when testing threshold invariance, their threshold parameter(s) can be either constrained to be equal or freely estimated. Given that

discrimination parameters and difficulty parameters are usually tested simultaneously within the IRT framework, to better correspond to measurement invariance tests in IRT, in the current study, equality constraints were placed on only loading parameters considered to have DIF or only threshold parameters considered to have DIF. Note that for conditions with polytomous data in which the threshold parameters had DIF, two of the noninvariant threshold parameters for each item needed to be constrained to be equal across groups for identification purpose; for conditions with binary data with noninvariant threshold parameters, these noninvariant parameters must be constrained to be equal in the analysis in order to identify the model.

Table 2

Manipulated Factors for the Generated Multiple-group Bifactor Models with DIF

Factors	DIF in general factor loadings	DIF in specific factor loadings	DIF in threshold parameters
Total sample size	600 or 1200	600 or 1200	600 or 1200
The number of categories per item	2, 3, or 5	2, 3, or 5	2, 3, or 5
Size of latent mean difference for the general factor	0, -0.1, or -0.2	0, -0.1, or -0.2	0, -0.1, or -0.2
Magnitude of DIF	-0.05, -0.10, or -0.15	-0.05, -0.10, or -0.15	0.05, 0.10, or 0.15
Data analysis procedure	Whether setting equality constraints on the general factor loadings with DIF	Whether setting equality constraints on the specific factor loadings with DIF	Whether setting equality constraints on the threshold parameters with DIF not necessary for identification

Data Generation and Data Analysis Procedures

R was used for data generation and Mplus 7 (Muthén & Muthén, 2012) was employed for data analysis; these were implemented with the *MplusAutomation* package in R. The data generation models were multiple-group GRMs whose parameters were transformed using Equation 16 from the corresponding loading parameters and threshold parameters of the multiple-group categorical CFA models in the varied research conditions presented above. For each condition, 1000 datasets with a given sample size were simulated from each population. Problematic simulations in which convergence was not achieved in a given number of iterations were discarded before summarizing the outcome so the number of simulations for each research condition may vary. To ensure independence across research conditions, a random seed was produced for each data generation condition using a random number generator. Note that for each data generation condition, there were multiple data analysis conditions.

To scale the latent factors, the means of all factors were fixed to 0 and the variances of all factors were fixed to 1 in the reference group, and loadings and threshold parameters (or intercepts) of at least one item for each factor were constrained to be invariant across groups. In the current study, all loadings and threshold parameters (or intercepts) were constrained to be equal for the conditions in which no DIF was generated. For the conditions in which DIF existed, the loadings and threshold parameters that were generated to be invariant were constrained to be equal across groups, and those with DIF were either estimated freely or constrained between groups, depending on data analysis conditions. Given that the latent mean differences in the general factor and specific factors are

indeterminate, they are unable to be estimated simultaneously in analysis. Given that specific factor mean differences were not the focus in the current study, when analyzing the latent mean difference of the general factor, the mean of the general factor was set to 0 in the reference group, the means of the specific factors were set to 0 in the both groups, and the mean of the general factor in the focal group was freely estimated.

In the analysis phase for multiple-group categorical CFA, DELTA parameterization in which the variances of latent response variates are set to 1 in the reference group was applied given that it performs better than THETA parameterization in some cases and there was no interest in testing invariance of the residual variance in the current study. As mentioned previously, in the data generation phase of the current study, variances of all latent response variates were generated to be 1 in both groups such that the constraints setting by DELTA parameterization would not influence the interpretation of the magnitude of the estimates for other parameters (e.g., latent mean difference of the general factor).

In addition to the constraints mentioned above, for multiple-group categorical CFA analysis based on binary data using misspecified unidimensional models, given that there is only one threshold parameter for each item, the variance of the latent continuous variate underlying item 1 (X_1^*) was fixed to 1 in the focal group. For multiple-group categorical CFA analysis based on polytomous data using bifactor models, the first two threshold parameters for each measured variable were constrained to be invariant across groups regardless of whether they had DIF; for binary data cases, the only threshold parameter of each item was constrained to be equal across groups regardless of whether DIF was

generated for them, and variances of all the continuous latent variates (X_1^*-X_{12}^*) were fixed to 1 in the focal group.

When applying multiple-group categorical CFA, the WLSMV estimator was applied and the DIFFTEST function in Mplus was used for testing latent mean difference of the general factor for each generated dataset. When conducting DIFFTEST, the more restricted model was the model constraining latent means of the general factor to be zero in both groups, and the less constrained model was the one freely estimating the latent mean of the general factor in the focal group. Other parameters of these two models were specified in the same way following the research conditions.

In the analysis phase for multiple-group traditional CFA, the ordered-categorical data were treated as continuous data and MLR estimator was applied. The Satorra-Bentler scaled chi-square difference test was used to test the latent mean difference of the general factor across groups. When conducting chi-square difference tests, the more restricted model was the model constraining latent means of the general factor to be zero in both of the two groups, and the less constrained model was the one freely estimating the latent mean of the general factor in the focal group. Other parameters of these two models were specified in the same way following the research conditions.

Outcomes of Interest

For a given research condition, the outcomes of interest were summarized across all replications with a proper solution. Results for replications with estimation problems were excluded from computation of the summary statistics. To obtain a comprehensive assessment of the impact of manipulated factors on the estimates and tests of differences in latent means between populations, the summarized outcomes for each condition

included Type I error rate or power, estimation bias, relative estimation bias, estimated variance, mean CFI, mean SRMR or WRMR, and mean RMSEA.

Type I Error Rates and Empirical Powers

For research conditions in which the effect size of the latent mean difference was generated to be zero, Type I error rate refers to the proportion of replications in which the null hypothesis that the latent mean of the general factor in the focal group is zero was rejected based on DIFFTEST or the Satorra-Bentler scaled chi-square difference test. For research conditions in which the effect size of the latent mean difference was non-zero in the population, power refers the proportion of replications in which the null hypothesis that the latent mean of the general factor in the focal group is zero was rejected based on DIFFTEST or the Satorra-Bentler scaled chi-square difference test. Empirical powers were interpreted for a condition only if the respective Type I error rate falls in the acceptable range of .025-.075, as designated by Bradley's liberal criterion (1978).

Estimation Biases, Relative Estimation Biases, and Variances for Latent Mean Difference

Estimation bias was the difference between the mean of estimated latent mean differences across replications within a condition and the population latent mean difference. The corresponding equation is:

$$\text{Estimation Bias} = \mathrm{E}(\hat{\theta}) - \theta \tag{21}$$

where $\mathrm{E}(\hat{\theta})$ denotes the average value of estimated latent mean differences across replications, and θ denotes the true value of the latent mean difference in the population. Relative estimation bias was the ratio of estimation bias to the true value of the latent mean difference in the population, and the corresponding equation is:

$$\text{Relative Estimation Bias} = (\text{E}(\hat{\theta}) - \theta) / \theta \quad (22)$$

Also, the variances of estimated latent mean differences across replications were computed.

Coverage Rates of 95% Confidence Interval

Coverage rate of the 95% confidence interval was the proportion of replications within a condition for which the population value of the difference in the general factor means falls within the computed 95% confidence interval for this difference in means.

Model Fit Indices

Model fit indices, including CFI, SRMR or WRMR, and RMSEA were collected because they are generally recommended for use in judging overall model fit. CFIs and RMSEAs were collected for all models, SRMRs were collected for models analyzed using the MLR estimator, and WRMRs were collected for models analyzed using WLSMV estimators. In the current study, these indices were examined to determine descriptively whether they were sensitive to misspecification of the noninvariant parameters.

Comparative Fit Index (CFI). CFI is an incremental fit index, which is defined as:

$$\text{CFI} = 1 - \left(\frac{\chi_T^2 - df_T}{\chi_N^2 - df_N}\right) \quad (23)$$

where χ_T^2 and χ_N^2 are chi-square statistics for the tested model and the null model in which only variances of observed variables are estimated, and df_T and df_N are the corresponding degree of freedom for these two models, respectively. Hu and Bentler (1999) suggested that a CFI values of .95 or higher indicates a good model fit. According to findings of Cheung and Rensvold (2002), a reduction of .01 or less in CFI suggests that hypothesis of invariance should not be rejected. CFIs obtained from replications with proper solutions were averaged for each condition to obtain the mean CFI.

Standardized Root Mean Square Residual (SRMR). SRMR is an absolute model fit index and reflects the mean of absolute correlation residuals. Hu and Bentler (1999) suggested that the acceptable value for SRMR should be equal or less than .08, and values of .05 indicate a good model fit (e.g., Kline, 2011). SRMRs obtained from replications with proper solutions were averaged for each condition to obtain the mean of SRMR.

Weighted Root Mean Square Residual (WRMR). Similar to SRMR, WRMR is also computed based on residuals, which was proposed by Muthén (1998-2004) for models using WLSMV estimators for ordered categorical data. As suggested by Yu & Muthén, (2002), values of 0.9 of WRMR suggest good fit, while Yu (2002) recommended a higher cutoff of 1.0 as a criterion for good fit. WRMRs obtained from replications with proper solutions were averaged for each condition to obtain the mean WRMR.

Root Mean Square Error of Approximation (RMSEA). RMSEA is a parsimony-corrected model fit index. The population RMSEA may be estimated based on fitting a hypothesized model to a sample as:

$$\text{RMSEA} = \sqrt{\frac{\chi^2 - df}{df\ (N-1)}} \tag{24}$$

where χ^2 is the chi-square statistic obtained when fitting a given model to a sample, df is the model degrees of freedom, and *N* is the sample size. As recommended by Browne and Cudeck (1993), values of less than .05 indicate a very good fit; values between .05 and .08 indicate a fair fit; and values larger than .10 indicate a bad fit. RMSEAs obtained from replications with proper solutions were averaged for each condition to obtain the mean RMSEA.

Chapter 3: Results

Factors Influencing the Latent Mean Comparisons for the General Factor in the No DIF Conditions

When exploring the impact of fitting bifactor ordered-categorical data with unidimensional models and treating ordered-categorical data as continuous data when evaluating latent mean comparisons for the general factor, all item parameters were generated to be invariant across groups and constrained to be equal in analysis such that there were no confounding effects due to DIF. The degree of unidimensionality varied in terms of different sizes of the explained common variance among all items due to the general factor (ECV) because the specific factor loadings varied while the general factor loadings remained the same for different data generation conditions. Also, sample sizes, effect sizes of the latent mean difference of the general factor, and the numbers of categories per item were also varied when generating the data. The generated bifactor polytomous data were analyzed using the following four strategies: unidimensional model with MLR estimator, unidimensional model with WLSMV estimator, bifactor model with MLR estimator, and bifactor model with WLSMV estimator. The generated bifactor binary datasets were analyzed using either unidimensional model with WLSMV estimator or bifactor model with WLSMV estimator.

Factors Influencing the Estimation Bias

The estimation bias of the general factor mean difference for the generated invariant bifactor ordered-categorical data is shown in Figure 2 and Tables B1-B6. Results indicated all the manipulated factors, including the degree of unidimensionality (sizes of specific factor loadings), sample size, the effect size of the latent mean difference for the general

factor, and the number of categories per item and estimation strategies applied in analysis, influenced the estimation bias of the general factor mean difference. Joint effects among these factors were also found.

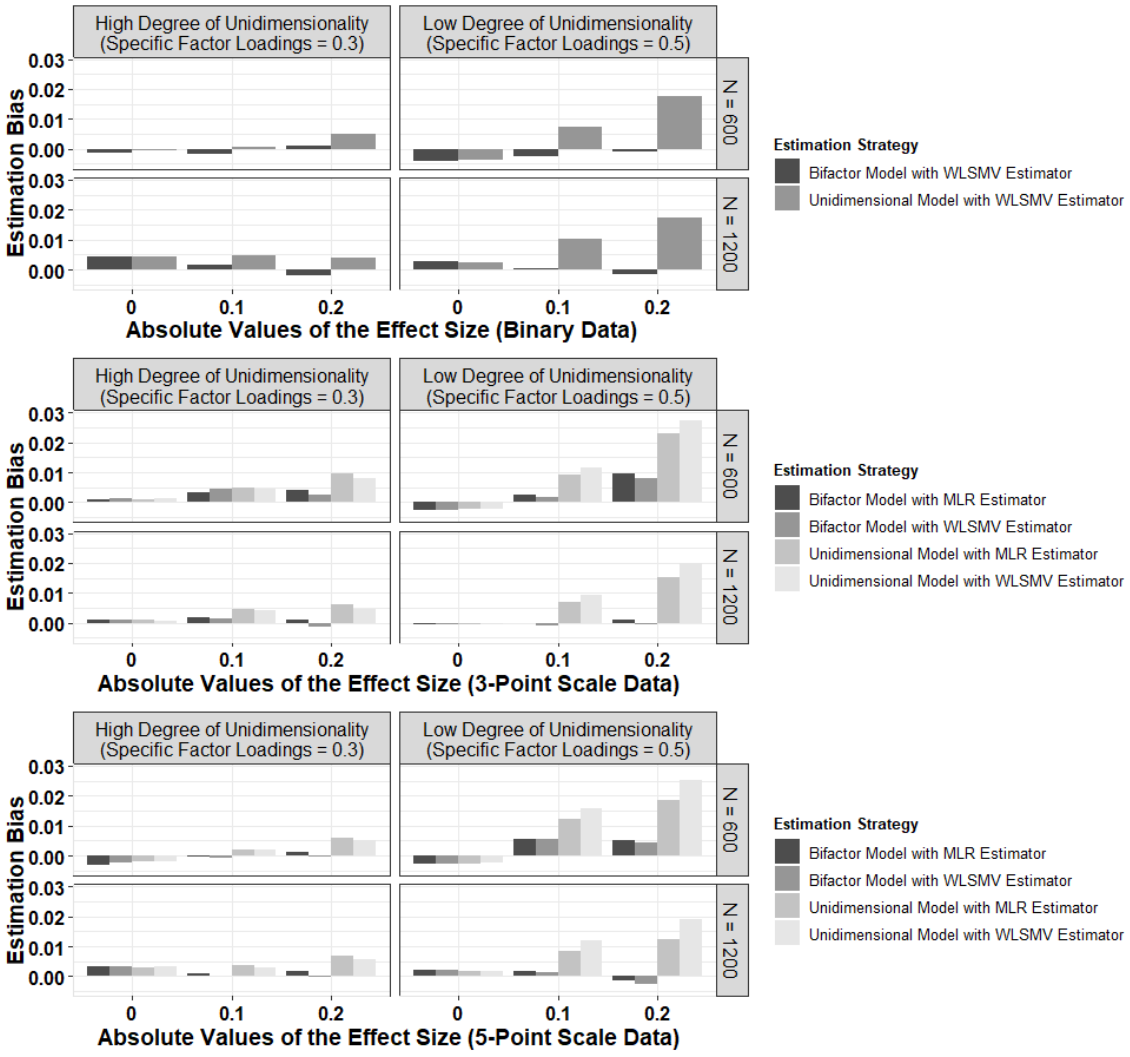

Figure 2 The Estimation Bias of the General Factor Mean Difference under No DIF Conditions

As shown in Figure 2 and Tables B1-B2, for the 3-point scale data, when the effect size of the latent mean difference for the general factor was generated to be zero and the

total sample size was 1200, the absolute values of estimation bias for the latent mean difference of the general factor ranged from .0005 to .0009 with a mean of .0007, which were ignorable and not influenced by other data generation conditions or data analysis conditions. For the conditions with effect size of zero, decreasing the total sample size from 1200 to 600 slightly increased the absolute values of estimation bias, making them ranged from .0008 to .0026 with a mean of .0017, which were not influenced by estimation strategies either.

For the 5-point scale data and the binary data, when the effect size of the general factor mean difference was generated to be zero, the absolute values of its estimation bias were a little higher than those for the 3-point scale data in general, and they were not influenced by neither estimation strategies nor the total sample size obviously (shown in Figure 2). When the effect size was zero, the absolute values of the estimation bias ranged from .0018 to .0034 with a mean of .0026 for the 5-point scale data (shown in Tables B1 and B2), and they ranged from .0003 to .0044 with a mean of .0029 for the binary data (shown in Tables B1 and B2).

As shown in Figure 2, for the generated invariant data, when the absolute values of effect size of the general factor mean difference increased, the absolute values of its estimation bias conditioning on other factors also increased in general, and the magnitude of the increase depended on total sample sizes and the joint factors of the degree of unidimensionality and analysis strategies.

To be specific, for the 3-point scale data, when the absolute value of the effect size was generated to be 0.1, the absolute values of estimation bias ranged from .0002 to .0094 and from .0019 to .0117 for the conditions with total sample size of 1200 and 600,

respectively; when the absolute value of the effect size was generated to be 0.2, the absolute values of estimation bias ranged from .0003 to .0198 and from .0025 to .0275 for the conditions with total sample size of 1200 and 600, respectively (shown in Tables B3-B6).

For the 5-point scale data, when the absolute value of the effect size was generated to be 0.1, the absolute values of estimation bias ranged from .0002 to .0118 and from .0002 to .0156 for the conditions with total sample size of 1200 and 600, respectively; when the absolute value of the effect size was generated to be 0.2, the absolute values of estimation bias ranged from .0001 to .0190 and from .0004 to .0251 for the conditions with total sample size of 1200 and 600, respectively (shown in Tables B3-B6).

Different from the data with 3 or 5 categories per item, for the binary data, decreasing the sample size did not increase the absolute values of the estimation bias when the effect size of the general factor mean difference was nonzero (shown in Figure 2). For the binary data, when the absolute value of the effect size was generated to be 0.1, the absolute values of estimation bias ranged from .0005 to .0101 and from .0007 to .0074 for the conditions with total sample size of 1200 and 600, respectively; when the absolute value of the effect size was generated to be 0.2, the absolute values of estimation bias ranged from .0017 to .0175 and from .0010 to .0176 for the conditions with total sample size of 1200 and 600, respectively (shown in Tables B3-B6).

In addition to the total sample size, the joint factors of the degree of unidimensionality and analysis strategies also had an influence on the estimation bias of the latent mean difference for the general factor when its effect size was generated to be nonzero. As indicated in Figure 2, for the data with 3 or 5 categories per item, when the total sample size was 1200, no matter what the effect size of the latent mean difference for

the general factor was, the absolute values of estimation bias was minimal as long as the data were fitted with bifactor models, and these values increased for some conditions when the total sample size decreased to 600. For the binary data, the absolute values of the estimation bias were also minimum if the data were fitted with bifactor models, and they were not influenced by the total sample size (shown in Figure 2). For the data with any number of categories (i.e., 2, 3 or 5), in comparison with the conditions in which the generated bifactor data was fitted with bifactor models, fitting the same generated data with unidimensional models produced more positive estimation bias in general (shown in Figures 1-3). The increase in the estimation bias was much more substantial for conditions involved the more multidimensional data (i.e., specific factor loadings = 0.5) than for conditions involved the less multidimensional data (i.e., specific factor loadings = 0.3), and it was also more substantial for the conditions with effect size for the general factor mean difference of -0.2 than for the conditions with the effect size of -0.1 (shown in Figure 2).

As shown in Figure 2, for the generated bifactor 3-point or 5-point scale data, the selection of estimators (i.e., WLSMV or MLR) had little influence on the estimation bias of the latent mean difference for the general factor as long as the data was fitted with bifactor models. When the more multidimensional data (i.e., specific factor loadings = 0.5) was fitted with unidimensional models, selecting the WLSMV estimator seemed to produce more estimation bias than selecting MLR estimator.

The relative estimation bias of the latent mean difference for the general factor is shown in Figure 3. The absolute values of the relative estimation bias were minimal when the generated bifactor data was fitted with bifactor models and the total sample size was 1200, and when decreasing the total sample size to 600, these values increased for the 3-

point scale and 5-point scale data and remained similar for the binary data. For the generated data with 2, 3, or 5 categories, falsely fitting the bifactor data with unidimensional models resulted in relative estimation bias with absolute values of around 2.5-5% for most conditions involving less multidimensional data (i.e., specific factor loadings = 0.3), and these values reached around 10-15% for the conditions involving more multidimensional data (i.e., specific factor loadings = 0.5). Different from the estimation bias (i.e., E ($\hat{\theta}$) – θ), the relative estimation bias (i.e., (E ($\hat{\theta}$) – θ) / θ) was not influenced substantially by the magnitude of the effect size when the effect size was generated to be nonzero.

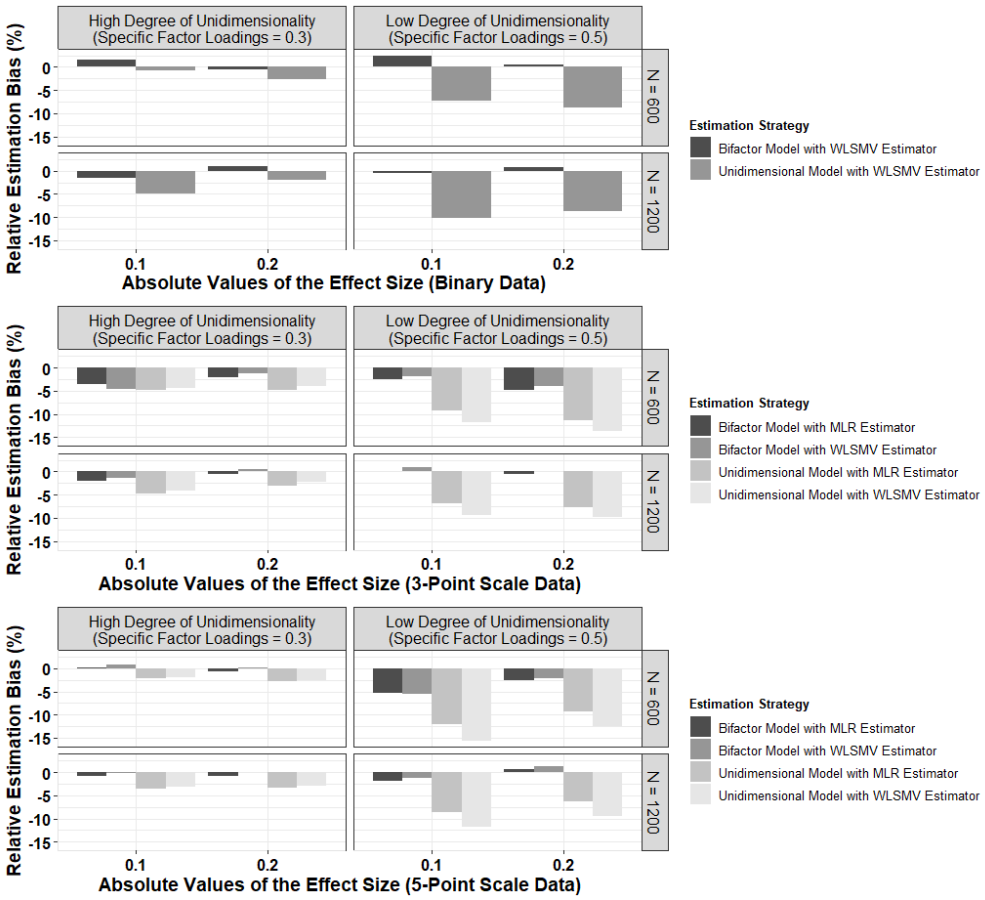

Figure 3 The Relative Estimation Bias of the General Factor Mean Difference under No DIF Conditions

Factors Influencing the Type I Error Rate/Power

The Type I error rate or power to detect the general factor mean difference for the generated invariant bifactor ordered-categorical data is shown in Figure 4 and Tables B1-B6.

Results regarding Type I error rates are shown in Tables B1 and B2. Type I error rates fell in the limits of .025 to .075 for all conditions involving invariant data. Also, for the generated data with 3 or 5 categories per item, Type I error rates obtained using the WLSMV estimator (ranged from .040 to .069) were a little higher than those obtained using the MLR estimator (ranged from .032 to .043) for a given generated data and analysis model (shown in Figures 4). When applying the WLSMV estimator, the relatively inflated Type I error rates (i.e., above .06) usually occurred when the total sample size was large or the datasets were binary. The selection of analysis model (unidimensional models vs. bifactor models) seemed to have no obvious influence on Type I error rates.

As shown in Figure 4, the most dominant factors influencing power were the effect size of the general factor mean difference and the total sample size. For the 3-point scale data, when the effect size was -0.1, the values of the power ranged from .170 to .206 and from .335 to .385 for the conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .547 to .607 and from .863 to .892 for the conditions with total sample size of 600 and 1200, respectively (shown in Tables B3-B6). For the 5-point scale data, when the effect size was -0.1, powers ranged from .165 to .229 and from .320 to .405 for the conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .564 to .672 and from .897 to .933 for conditions with total sample size of 600 and 1200, respectively (shown in Tables

B3-B6). For binary data, when the effect size was -0.1, powers ranged from .210 to .232 and from .358 to .378 for conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .597 to .646 and from .886 to .895 for conditions with total sample size of 600 and 1200, respectively (shown in Tables B3-B6).

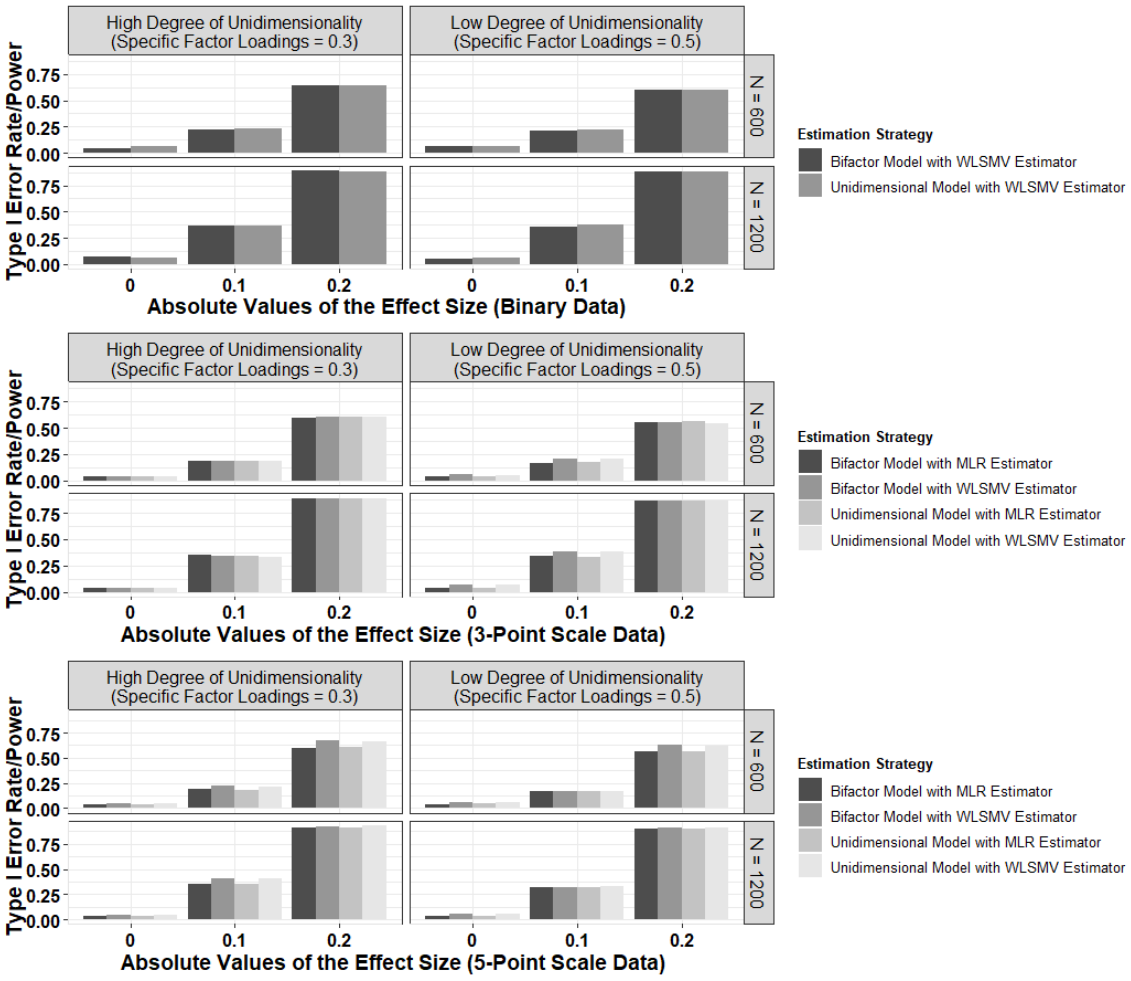

Figure 4 Empirical Type I Error Rates/Powers to Detect the General Factor Mean Difference under No DIF Conditions

In addition to the effect size of the general factor mean difference and the total sample size, the selection of the estimator might also influence the power to detect the

general factor mean difference. As shown in Figure 4, for a given generated data set and analysis model (unidimensional model or bifactor model), the power obtained using the WLSMV estimator seemed to be higher than that obtained using the MLR estimator in most of the cases.

For a given total sample size and effect size of the general factor mean difference, the variation of the values in power among conditions using MLR estimator seemed to be smaller than that using the WLSMV estimator. For the MLR estimator, when the effect size was -0.1, powers ranged from .165 to .188 and from .320 to .355 for the conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .549 to .612 and from .863 to .908 for the conditions with total sample size of 600 and 1200, respectively (shown in Tables B3-B6). For the WLSMV estimator, when the effect size was -0.1, powers ranged from .166 to .232 and from .325 to .405 for the conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .547 to .672 and from .864 to .933 for the conditions with total sample size of 600 and 1200, respectively (shown in Tables B3-B6).

As shown in Tables B3-B6, a part of the large variation in the values of power obtained using the WLSMV estimator among the conditions for a given sample size and effect size can be explained by the data generation factors of the number of categories per item and the degree of unidimensionality in the current research scenario. For example, in most cases, conditioning on other factors, the powers to detect the general factor mean difference obtained using the WLSMV estimator were largest for the 5-point scale data, and they were slightly higher for the conditions involving less multidimensional data than those for the conditions involving more multidimensional data.

As shown in Figures 2-4, for the same generated bifactor data, although fitting them with unidimensional models produced varying degrees of estimation bias in the latent mean difference of the general factor, power was not influenced by the analysis model applied (i.e., unidimensional models or bifactor models) using a given estimator (i.e., the MLR estimator or the WLSMV estimator).

Estimated Variance

In the No DIF conditions, the main factor that influenced the estimated variance of the latent mean difference in the general factor was sample size (shown in Tables B1-B6). When the total sample size was 600, the estimated variances for these No DIF conditions ranged from .006 to .009; when the total sample size was 1200, they were .003 or .004. In addition to the total sample size, applying the unidimensional model usually led to a smaller estimated variance in comparison with applying the bifactor model, and the difference in the estimated variance for a given generated dataset was .001 or .002. Also, in most cases, the estimated variance for the general factor mean difference tended to be smaller for polytomous data than that for binary data for a given total sample size. Given the level of precision reported (rounding to the thousandths place), not all differences in estimated variances were reported. The selection of the estimator (i.e., the MLR estimator or the WLSMV estimator), the effect size of the general factor mean difference, and the degree of unidimensionality for the generated bifactor data had no obvious impact on the estimated variance for the general factor mean difference.

Coverage Rates of the 95% Confidence Interval

In the No DIF conditions, almost all the coverage rates of the 95% confidence interval were above .950. Only for one condition, the coverage rate dropped to .948 because

of relatively serious estimation bias (i.e., .0251 in the absolute value) resulting from fitting the generated bifactor data using a unidimensional model when the total sample size was 600 and the effect size of the general factor mean difference was generated to be -0.2.

Goodness of Fit Indices

The means of the goodness of fit indices (i.e., CFI, SRMR/WRMR, and RMSEA) for the No DIF conditions are shown in Tables B7-B12.

Means of CFIs. As shown in Tables B7-B12, the means of CFIs for most of the conditions in which the generated bifactor ordinal data were analyzed with bifactor models using the WLSMV estimator were .999 when the general factor mean difference was freely estimated, and in a few small sample size conditions, they were .998. When the effect size of the general factor mean difference was 0, constraining the latent mean difference in the general factor to be 0 did not result in any decrease in these means of CFIs, and it even increased the means of CFIs for a few conditions. When the effect size of the general factor mean difference was -0.1, the drop of the means of CFIs due to constraining the latent mean difference of the general factor to be zero were 0.001 for almost all the conditions except for one condition in which the mean of the CFI did not change based on the level of precision reported (rounding to the thousandths place). When the effect size of the general factor mean difference was -0.2, the conditions involving less multidimensional data evidenced drops in the means of CFIs due to constraining to zero the latent mean difference of the general factor of 0.005 and .007 for binary data and polytomous data, respectively; in the conditions involving more multidimensional data, reductions in the means of CFIs were .003 and .004 for binary data and polytomous data, respectively.

The means of CFIs for the conditions in which the less multidimensional data (i.e., specific factor loadings = 0.3) was analyzed with unidimensional models using the WLSMV estimator ranged from .977 to .984 (shown in Tables B7, B9, and B11). For the same generated dataset and the same estimator (i.e., the WLSMV estimator), in comparison with the bifactor analysis model, the drop of the means of CFIs resulting from fitting them with the unidimensional model were from .015 to .016 and from .019 to .021 for binary data and polytomous data, respectively. For these misspecified conditions with unidimensional analysis models, constraining the general factor mean difference to be zero increased the means of CFIs when the effect size was generated to be 0; even when the effect size of the general factor mean difference was generated to be -0.1, all but one of the means of CFIs still increased after constraining the general factor mean difference to be zero. Also, the increase in the means of CFIs due to constraining the general factor mean difference to be zero was especially obvious for polytomous data. When the effect size of the general factor mean difference was -0.2, imposing the equality constraint on the general factor mean difference in unidimensional models resulted in the drop of .004 and .001 for the means of CFIs for binary data and 3-point scale data, respectively. But for the 5-point scale data, they still increased even when constraining the general factor mean difference of -0.2 to be equal.

With respect to the generated more multidimensional data (i.e., specific factor loadings = 0.5), the means of CFIs for the conditions with unidimensional analysis models using the WLSMV estimator ranged from .887 to .912(shown in Tables B8, B10 and B12). For binary, bifactor data analyzed with the WLSMV estimator, the means of CFIs were smaller by .087 and .092 with small sample size and large sample size, respectively, when

fitting them with the misspecified unidimensional model in comparison with the bifactor analysis model. For the 3-point scale data, the drops were .101 and .105-.106 for small sample size conditions and large sample size conditions, respectively; for the 5-point scale data, they were .107-.108 and .112 for small and large sample size conditions, respectively. After incorrectly fitting the more multidimensional data with unidimensional models using the WLSMV estimator, imposing the equality constraint on the general factor mean difference increased the means of CFIs no matter what the effect size was generated to be (i.e., 0, -0.1 or -0.2), and such increase got large as the number of categories per item became more.

For the conditions in which the generated ordinal data were treated as continuous and the MLR estimator was applied, the means of CFIs for bifactor analysis conditions ranged from .996 to .999, which was slightly influenced by the data generation conditions (i.e., the total sample size, the number of categories per item, and the degree of unidimensionality). After constraining the general factor mean difference to be zero, there was no change in the means of CFIs for all conditions with the effect size of 0 and most conditions with the effect size of -0.1; the drop was .001 for one condition with the effect size of -0.1 and all conditions with the effect size of -0.2 (shown in Tables B7-B12).

The means of CFIs for the conditions in which the generated less multidimensional data (i.e., specific factor loadings = 0.3) was analyzed with unidimensional models using the MLR estimator ranged from .946 to .953(shown in Tables B7, B9, and B11). For the same generated dataset and the same estimator (i.e., the MLR estimator), in comparison with the bifactor analysis model, the drop of the means of CFIs resulting from fitting them with the unidimensional model were around .045 and .050 for 3- and 5-point scale data,

respectively. For these unidimensional analysis models using the MLR estimator, there were no changes in the means of CFIs after constraining the general factor mean difference to be zero for all conditions with the effect size of 0 and half of the conditions with the effect size of -0.1, and they dropped by .001 and .002 for the other half of conditions with the effect size of -0.1 and all conditions with the effect size of -0.2, respectively.

With respect to the generated more multidimensional data (i.e., specific factor loadings = 0.5), the means of CFIs for the unidimensional analysis models using the MLR estimator ranged from .727 to .751 (shown in Tables B8, B10, and B12). To be specific, in comparison with the bifactor analysis model using the MLR estimator, the decrease of the means in CFIs due to fitting them with the unidimensional model using the MLR estimator were 246-.247 and .248 - .249 for the 3-point scale data with small and large sample size, respectively, and they were around .270 for the 5-point scale data. For these unidimensional models with MLR estimator, imposing the equality constraint on the general factor mean difference, the means of CFIs did not drop or dropped by .001 when the effect size was 0 or -0.1, and they dropped by .001 or .002 when the effect size was -0.2.

Means of WRMRs. The means of WRMRs were reported for the conditions analyzed with the WLSMV estimator. As shown in Tables B7-B12, when the generated bifactor ordinal data were analyzed with bifactor models, the means of WRMRs ranged from .802 to 1.023. The means of WRMRs were influenced by the data generation conditions including the number of categories, total sample size, and degree of unidimensionality. For example, for binary data, means of WRMRs ranged from .969 to 1.023, while for polytomous data, they ranged from .802 to .857. Also, larger sample size

and high degree of unidimensionality were associated with larger values in the means of WRMRs.

After imposing equality constraints on the general factor mean difference generated to be 0 in the bifactor analysis models with the WLSMV estimator, the means of WRMRs ranged from 1.003 to 1.050 and from .854 to .911 for binary data and polytomous data, respectively. When the general factor mean difference with effect size of -0.1 was constrained to be equal, the means of WRMRs ranged from 1.044 to 1.113 and from .910 to 1.070 for binary data and polytomous data, respectively. When the effect size of the constrained general factor mean difference was -0.2, the means of WRMRs ranged from 1.129 to 1.292 and from 1.055 to 1.442 for binary data and polytomous data, respectively. The changes of the means of WRMRs due to imposing equality constraints on the nonzero general factor mean difference were more substantial for polytomous data than those for binary data, with maximum of around 0.2 and 0.6 for the conditions with effect size of -0.1 and -0.2, respectively (shown in Tables B7-B12).

When the generated less multidimensional data (i.e., specific factor loadings = 0.3) were analyzed with unidimensional models using the WLSMV estimator, the means of WRMRs ranged from 1.225 to 1.442 and from 1.290 to 1.599 for binary and polytomous data, respectively. With respect to the generated more multidimensional data (i.e., specific factor loadings = 0.5), when fitting them with the misspecified unidimensional model using the WLSMV estimator, the means of WRMRs ranged from 2.639 to 3.599 and from 3.109 to 4.517 for binary and polytomous data, respectively. Larger number of categories per item and the larger sample size were associated with more increase in the means of WRMRs due to fitting the bifactor data using the unidimensional model (shown in Tables

B7-12). In the unidimensional analysis models, after imposing equality constraints on the general factor mean difference, the means of WRMRs became even larger but the increase was very small relative to that resulting from falsely fitting the bifactor data using the unidimensional model.

Means of SRMRs. As shown in Tables B7-B12, when the generated bifactor ordinal data was treated as continuous and analyzed using the MLR estimator, the means of SRMRs were reported. Using bifactor analysis models, the means of SRMRs ranged from .029 to .033 and from .021 to .024 for the small (Total N = 600) and large (i.e., Total N = 1200) sample conditions. In addition to the sample size, the means of SRMRs were also influenced by other data generation conditions slightly, such as the number of categories per item and the degree of unidimensionality. After imposing equality constraints on the general factor mean difference, the increase in the means of SRMRs ranged from .001 to .002, from .003 to .004, and from .008 to .012 for the conditions with effect sizes of 0, -0.1, and -0.2, respectively.

The means of SRMRs for the conditions in which the less multidimensional data (i.e., specific factor loadings = 0.3) was analyzed with unidimensional models was .046 and .039 - .040 when the total sample size was 600 and 1200, respectively (shown in Tables B7, B9, and B11). After imposing equality constraints on the general factor mean difference, the increase in the means of SRMRs was 0-.001, .002-.003, and .006-.007 for the conditions with the effect size of 0, -0.1 and -0.2, respectively.

With respect to the more multidimensional data (i.e., specific factor loadings = 0.5), fitting them using unidimensional models, the means of SRMRs became .091 and .096 for the 3- and 5-point scale data, respectively, when the total sample size was 600, and they

were .088 and .093 for the 3- and 5-point scale data when the total sample size was 1200 (shown in Tables B8, B10, and B12). The increase in the means of SRMRs resulting from imposing equality constraints on the general factor mean difference in these unidimensional models was 0, .001, and .003 for conditions with effect sizes of 0, -0.1, and -0.2, respectively.

Means of RMSEAs. As shown in Tables B7-B12, when the generated bifactor ordinal datasets were analyzed using bifactor models, the means of RMSEAs ranged from .006 to .013 with their magnitude influenced by data generation conditions including the total sample size, the number of categories per item, and the degree of unidimensionality. The selection of the estimator (i.e., the MLR estimator or the WLSMV estimator) seemed to have little impact on the means of RMSEAs when the model was correctly specified. After imposing equality constraints on the general factor mean difference, the means of RMSEAs did not change or even decreased when the effect size was 0. When the effect size was -0.1, setting equality constraints on the latent mean difference of the general factor yielded increases in means of RMSEAs ranging from .003 to .006 for analysis conditions with the WLSMV estimator, and 0 to .001 for conditions with the MLR estimator. When the effect size was -0.2, the increase in the means of RMSEAs ranged from .013 to .025 and from .002 to .005 for conditions with the WLSMV estimator and the MLR estimator, respectively, after imposing equality constraints on the general factor mean difference.

When generated less multidimensional data (i.e., specific factor loadings = 0.3) were fitted with unidimensional models, the means of RMSEAs ranged from .042 to .062, and the increase relative to the corresponding bifactor model using the same estimator

ranged from .034 to .053. With respect to the more muldata (i.e., specific factor loadings = 0.5), fitting them with unidimensional models, the means of RMSEAs ranged from .122 to .169 with the increase in comparison with the corresponding correctly specified model using the same estimator ranged from .112 to .160. After imposing equality constraints on the general factor mean difference in these unidimensional models, the means of RMSEAs did not change or even decreased for all conditions with the effect size of 0 and -0.1 and most conditions with the effect size of -0.2. Only for some conditions in which the data was generated with high degree of unidimensionality and the effect size of -0.2, the means of RMSEAs increased by .001.

Factors Influencing the Latent Mean Comparisons for the General Factor in the Conditions with DIF

In order to examine the influence of the DIF on latent mean comparisons of the general factor within the generated multiple-group ordinal bifactor datasets, the manipulated data generation conditions included the total sample size (i.e., 600 or 1200), the effect size of the general factor mean difference (i.e., 0, -0.1, or -0.2), the number of categories per item (2, 3, or 5), the type of parameters with DIF (general factor loadings, specific factor loadings, or threshold parameters), and the magnitude of DIF (i.e., -0.05, -0.10, or -0.15 for factor loadings; 0.05, 0.10, or 0.15 for threshold parameters). When analyzing the data, all generated datasets were analyzed using bifactor models with the WLSMV estimator, and the parameters with DIF were either freely estimated or constrained to be equal across groups. For conditions with DIF, all item parameters (i.e., general factor loadings, specific factor loadings, and threshold parameters) in the reference group and variance-covariance matrix and mean vectors for the latent factors were the same

as those for the No DIF conditions with low degree of unidimensionality (i.e., strong specific factor loadings) except that the DIF was present for different types of item parameters, so the subset of the No DIF conditions in which the data were generated with strong specific factor loadings and analyzed with bifactor models using the WLSMV estimator serve as the baseline conditions for evaluating the DIF conditions.

Factors Influencing the Estimation Bias

Estimation bias for the conditions with DIF in general factor loadings are shown in Figure 5 and Tables B13-B18. In comparison with the corresponding baseline conditions in which no DIF was generated and all parameters were constrained to be equal, the DIF in the general factor loadings seemed to have no influence on the estimation bias for the general factor mean difference if the general factor loadings with DIF were freely estimated (see Figure 5). As shown in Tables B2, B4, and B6, the estimation bias of the general factor mean difference for all baseline conditions ranged from -.0041 to .0081 with a mean of .0004. As shown in Figure 2, the estimation bias of these baseline conditions was slightly influenced by the total sample size and effect size of the general factor mean difference for the 3- and 5-point scale data. For the conditions with DIF in the general factor loadings, the estimation bias of the general factor mean difference ranged from -.0044 to .0049 with a mean of -.0001 when the model was correctly specified (shown in Tables B13-B18), which fell in the similar range as the baseline conditions. Unlike the baseline conditions, the estimation bias of these correctly specified models for the conditions with DIF in the general factor loadings was not influenced by the total sample size or the effect size of the general factor mean difference (shown in Figure 5).

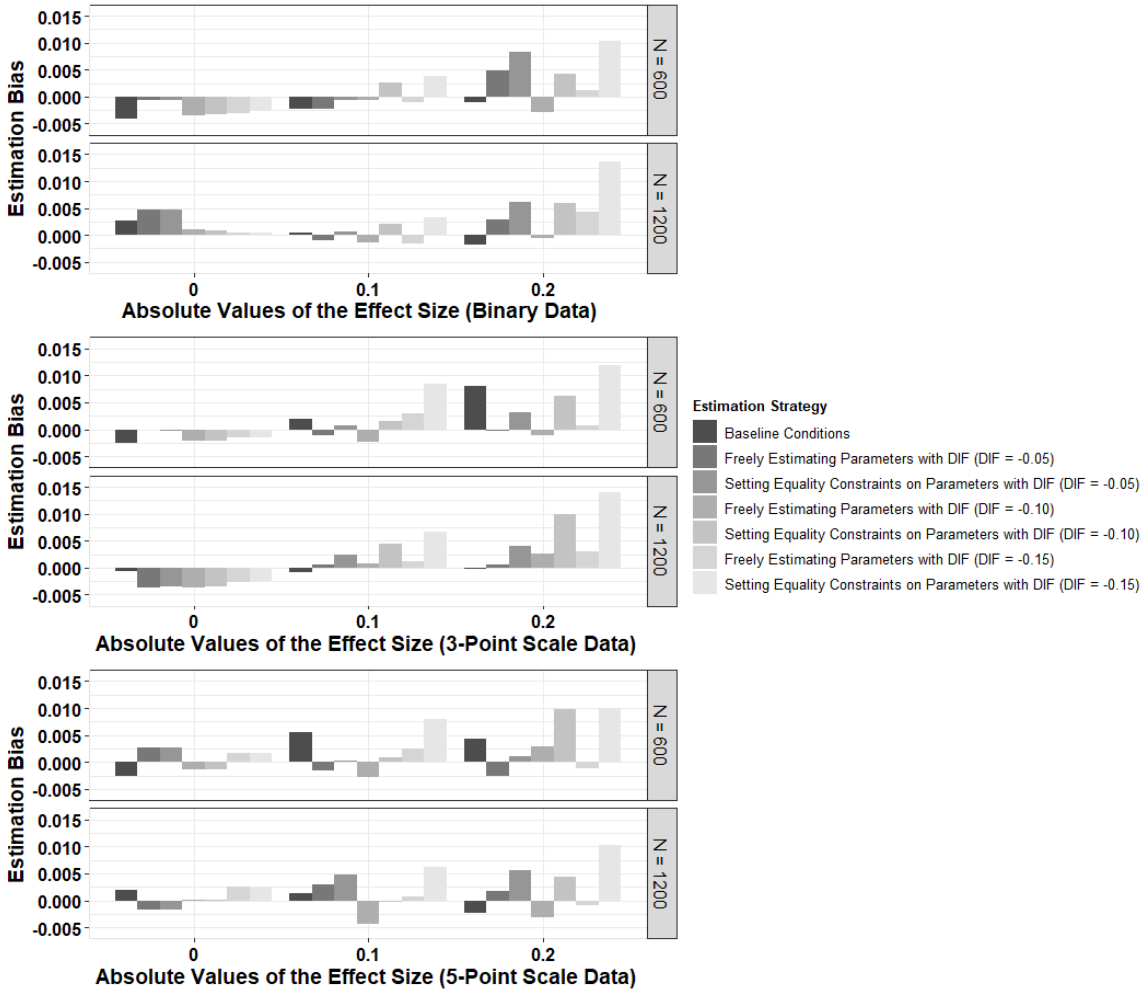

Figure 5 Estimation Bias of the General Factor Mean in the Conditions with DIF in General Factor Loadings

As shown in Tables B13-B18, when the magnitude of DIF in the general factor loadings were -0.05, the differences of the estimation bias for the general factor mean difference between the conditions with general factor loadings having DIF constrained to be equal and the corresponding correctly specified conditions ranged from -.0001 to .0002, from .0016 to .0018, and from .0034 to .0037 when the effect size was 0, -0.1, and -0.2, respectively. As the magnitude of DIF in the general factor loadings increased, the differences in estimation bias in the general factor mean difference between these two

analysis conditions remained similar for the conditions with the effect size of zero (i.e., ranged from -.0001 to .0002 and from 0 to .0003 when DIF = -0.10 and -0.15, respectively), and they increased obviously for the conditions with the effect size of -0.10 and -0.15. To be specific, when the DIF = -0.10, they ranged from .0033 to .0040 and from .0066 to .0075 for conditions with effect sizes for the general factor mean difference of -0.1 and -0.2, respectively. When the DIF = -0.15, they were .0048 and .0055 for binary and polytomous data, respectively, in the conditions with the effect size of -0.1, and they were .0091- .0093 and .0111- .0112 for binary and polytomous data in conditions with the effect size of -0.2.

In summary, as shown in Figure 5 and Tables B13-B18, for a given generated dataset with DIF in the general factor loadings, when constraining the general factor loadings with DIF to be equal across groups, in comparison with the correctly specified model with the general factor loadings with DIF freely estimated, there was no obvious change in the estimation bias for the general factor mean difference for the conditions with the effect size of 0, and there was substantial increase in the estimation bias for the general factor mean difference for conditions with nonzero effect size (i.e., -0.1 or -0.2). The magnitude of the increase in the estimation bias for the general factor mean difference resulting from constraining the general factor loadings with DIF to be equal across groups was mainly determined by the magnitude of DIF and the effect size of the general factor mean difference, and it was slightly influenced by the number of categories per item. Controlling for the magnitude of the DIF and the number of categories per item, when the effect size of the general factor mean difference was -0.2, the increase in its estimation bias due to setting equality constraints on the general factor loadings having DIF was around two times that for the conditions with the effect size of -0.1.

In addition to the estimation bias (i.e., E ($\hat{\theta}$) – θ), the relative estimation bias (i.e., (E ($\hat{\theta}$) – θ) / θ) was also reported in the current study. As shown in Figure 6, in the baseline conditions and the conditions with the general factor loadings having DIF freely estimated, most of the relative estimation biases were around 0, and their maximum values could reach around 5% in absolute values. In comparison with the correctly specified model, when the DIF = -0.05, constraining the general factor loadings with DIF to be equal across groups resulted in around 1.6-1.8% decrease in the relative estimation bias for the general factor mean difference estimates; when the DIF = -0.10, the decrease was around 3.2-3.9%; when the DIF = -0.15, the decrease was around 4.6-4.8% and 5.5-5.6% for binary data and polytomous data, respectively. As indicated in Figure 6, the changes in the relative estimation bias for the general factor mean difference estimates due to setting equality constraints on general factor loadings with DIF were not substantial relative to the estimation bias of these estimates for the correctly specified models when the DIF = -0.05 and -0.10, and only when DIF = -0.15, the relative estimation bias for the general factor mean difference estimates resulting from constraining the general factor loadings with DIF to be zero might need more attention. Unlike how the estimation bias of the general factor mean difference was influenced by its effect size, the relative estimation bias was not influenced by the effect size.

The results of (relative) estimation bias for the conditions with DIF in specific factor loadings are shown in Figures 7 and 8 and Tables B19-B24. Similar to conditions with DIF in general factor loadings, the DIF in specific factor loadings had no impact on the estimation bias for the general factor mean difference as long as the model was correctly specified. When the specific factor loadings with DIF were freely estimated, the estimation

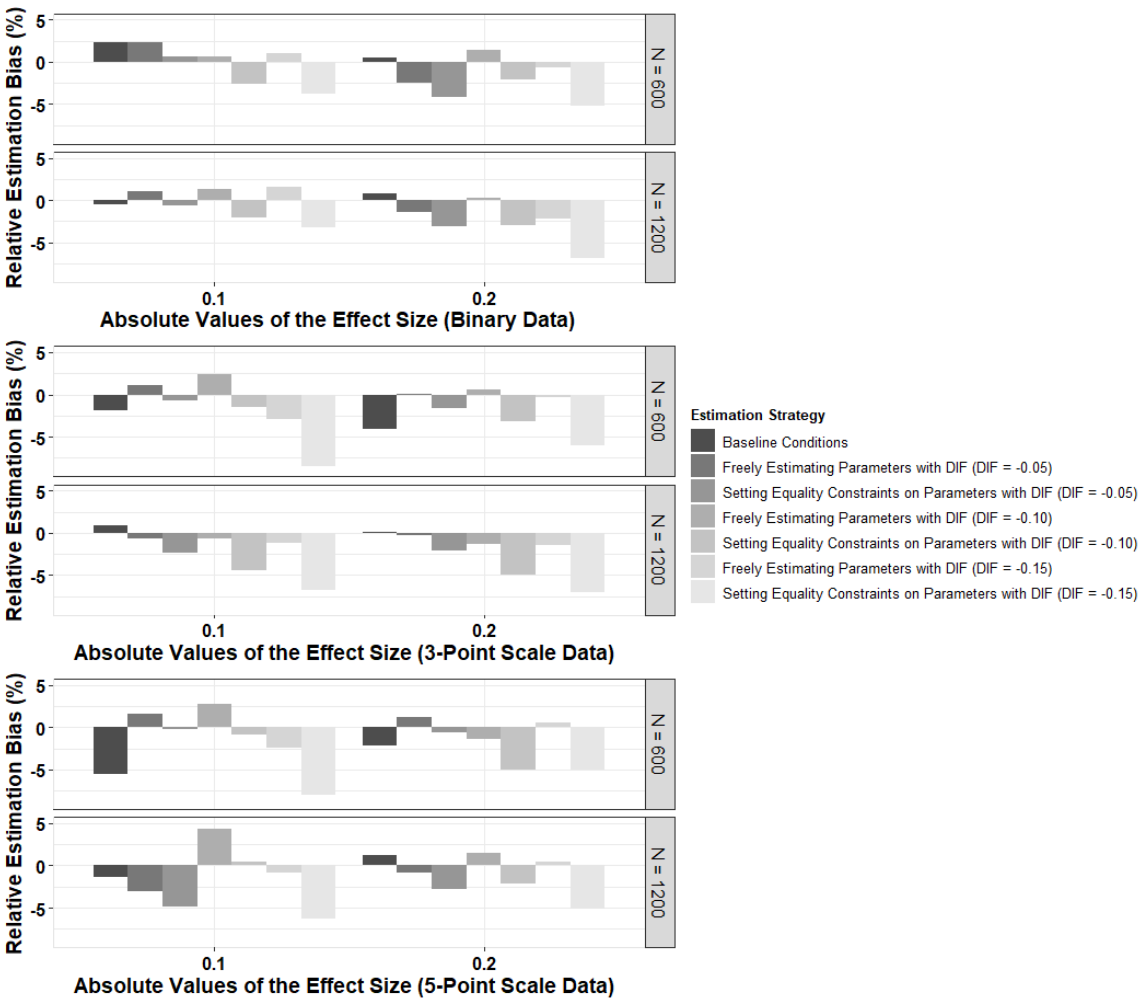

Figure 6 Relative Estimation Bias of the General Factor Mean in the Conditions with DIF in General Factor Loadings

bias ranged from -.0058 to .0068 with a mean of .0003, which fell in a similar range to the baseline conditions and the correctly specified conditions with DIF in the general factor loadings, and their absolute values slightly increased as the total sample size decreased from 1200 to 600.

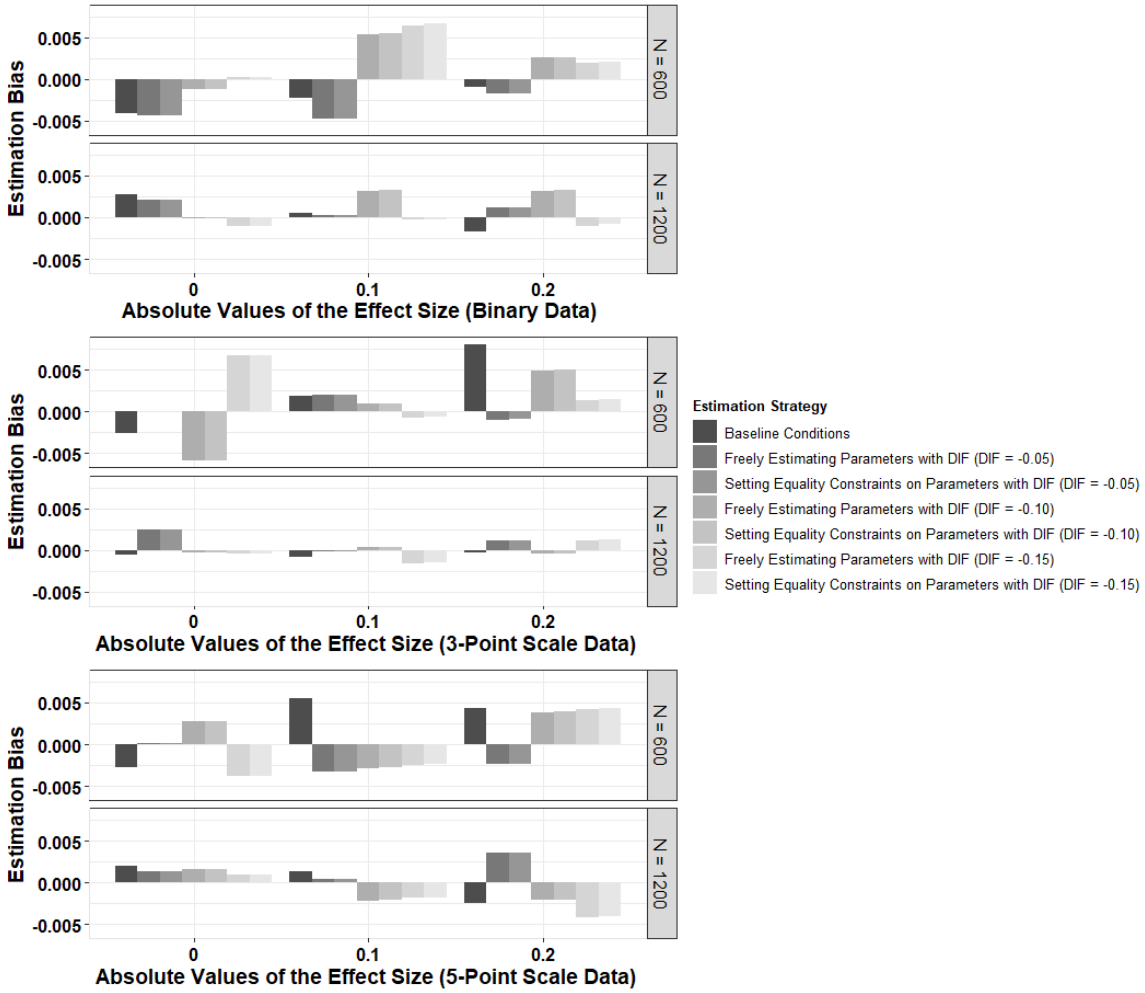

Figure 7 Estimation Bias of the General Factor Mean in the Conditions with DIF in Specific Factor Loadings

After setting equality constraints on the specific factor loadings with DIF, in comparison with the corresponding correctly specified model, the estimation bias of the general factor mean difference did not change for almost all conditions when the effect size was zero. When the effect size of the general factor mean difference was nonzero, the changes in estimation bias due to setting equality constraints on specific factor loadings with DIF were 0, .0001, or .0002 regardless of the size of the DIF (shown in Tables B19-B24).

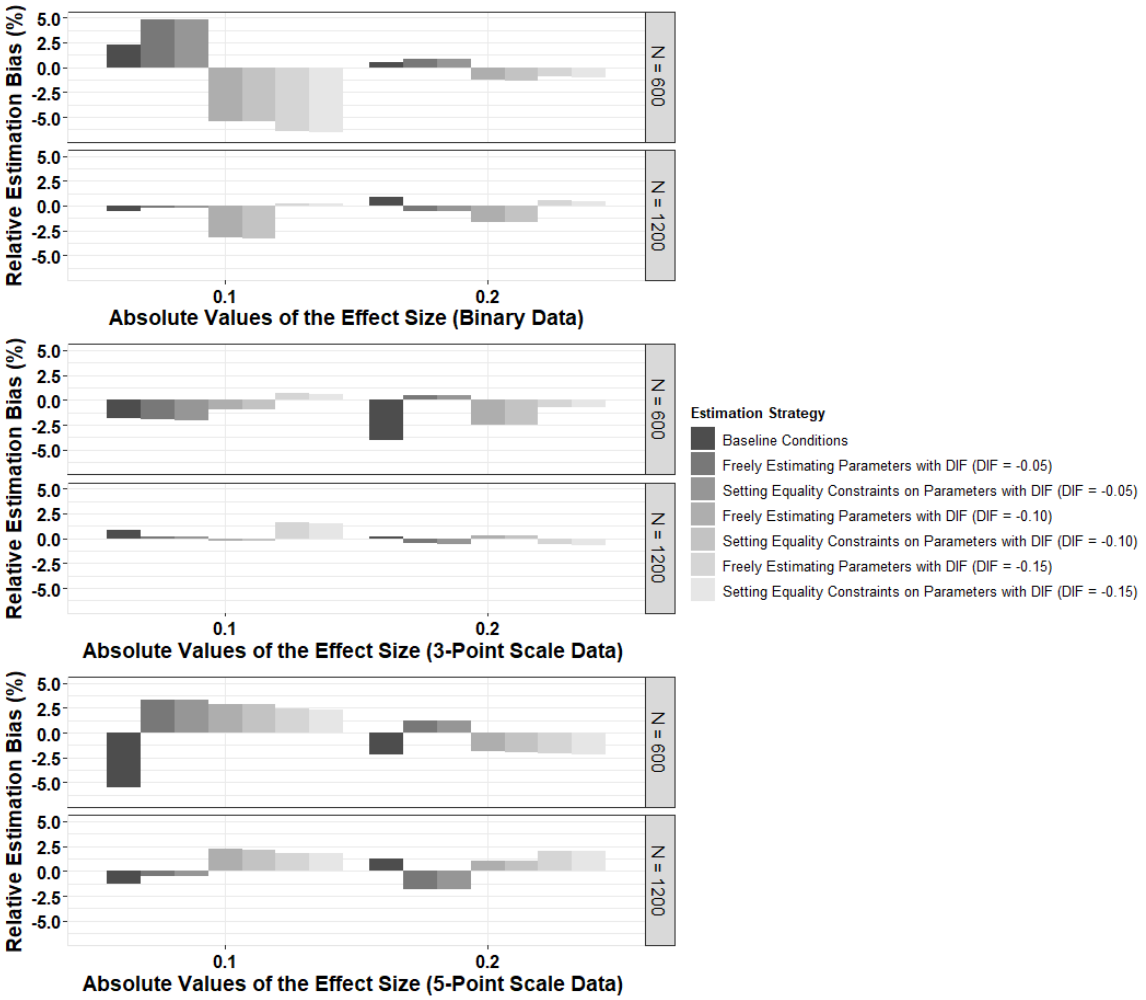

Figure 8 Relative Estimation Bias of the General Factor Mean in the Conditions with DIF in Specific Factor Loadings

As indicated by Figures 7 and 8, the changes in estimation bias or relative estimation bias resulting from setting equality constraints on the specific factor loadings with DIF were minimal relative to the estimation bias or the relative estimation bias for the correctly specified models.

The results of estimation bias for the conditions with DIF in threshold parameters are shown in Figure 9 and Tables B25-B30. When the DIF was present in threshold parameters, a constant of 0.05, 0.10, or 0.15 was added to all the threshold parameter(s) of

the noninvariant items. For binary data and the 3-point scale data, the noninvariant threshold parameters had to be constrained to be equal for identification purpose. For the 5-point scale data, the first two noninvariant threshold parameters for each noninvariant item were constrained to be equal for identification purposes and the other two noninvariant threshold parameters could be freely estimated. As shown in Tables B25-B30, for binary data and the 3-point scale data, the estimation bias of the general factor mean difference ranged from -.0199 to -.0123, from -.0381 to -.0286, and from -.0584 to -.0499 when the DIF in the threshold parameters was 0.05, 0.10, and 0.15, respectively. For the 5-point scale data, when two of the noninvariant threshold parameters for each item with DIF were freely estimated, the estimation bias of the general factor mean difference ranged from -.0131 to -.0099, from -.0252 to -.0203, and from -.0356 to -.0277 when the DIF in the threshold parameters was 0.05, 0.10, and 0.15, respectively. After constraining all threshold parameters with DIF to be equal for the 5-point scale data, the estimation bias of the general factor mean difference ranged from -.0204 to -.0180, from -.0403 to -.0343, and from -.0549 to .0495 when the DIF = 0.05, 0.10, and 0.15, respectively. For a given generated 5-point scale data with the DIF of 0.05 in threshold parameters, the changes in estimation bias for the general factor mean difference due to constraining more noninvariant threshold parameters to be equal were -.086 or -.084, -.074 or -.075, -.069 or -.067 when the effect size was 0, -0.1, and -0.2, respectively. When DIF was 0.10, these changes were -.0172 or -.0169, -.0153 or -.0151, -.0132 or -.0131 for the conditions with effect sizes of 0, -0.1, and -0.2, respectively; when the DIF was 0.15, they were -.0246 or -.0248, -.0218, -.0193 or -.0189 for the conditions with the effect size of 0, -0.1, and -0.2, respectively.

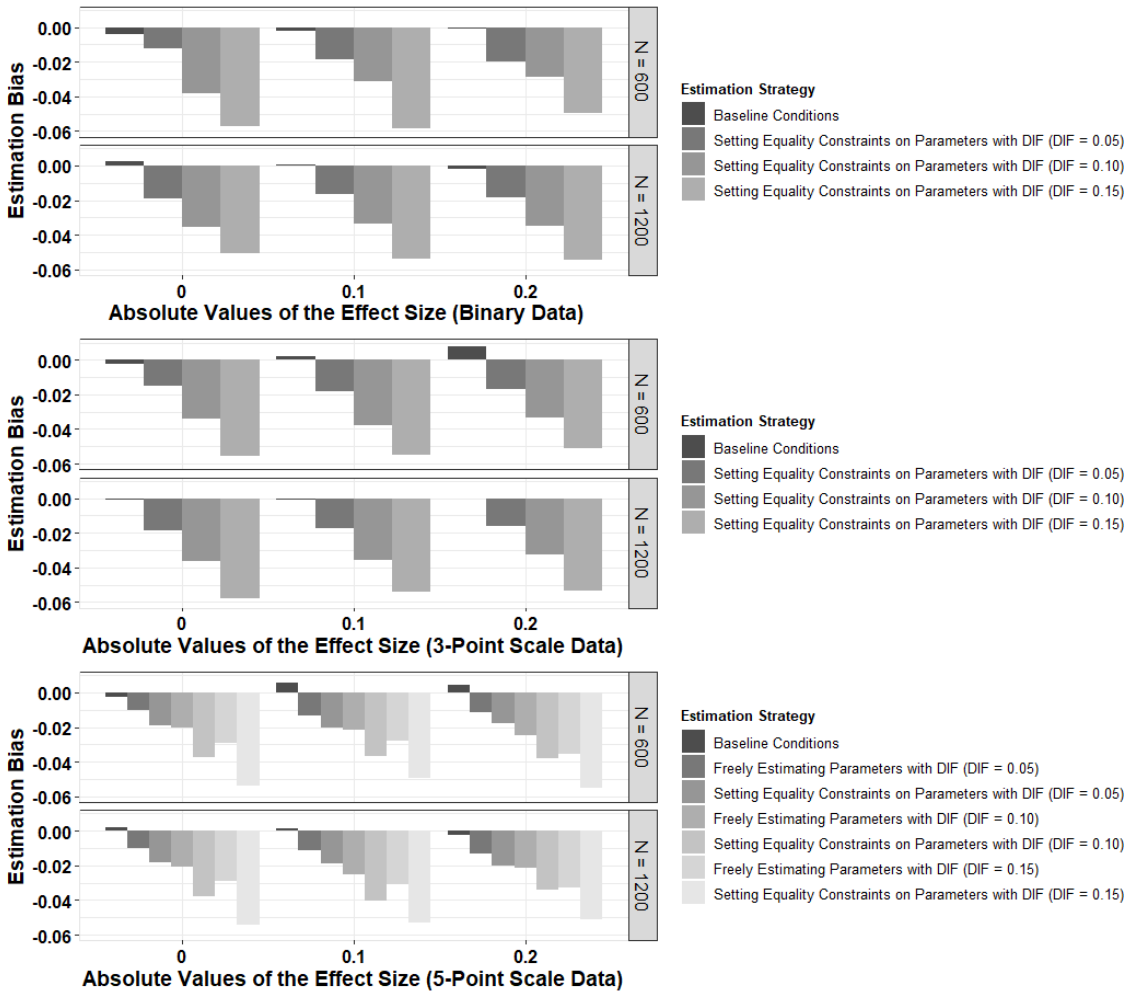

Figure 9 Estimation Bias of the General Factor Mean in the Conditions with DIF in Threshold Parameters

As indicated in Figure 9, the estimation biases of the general factor mean difference for the conditions with noninvariant threshold parameters constrained to be equal were negative and very substantial relative to the corresponding baseline conditions, and the main factor that influenced the estimation bias was the magnitude of DIF in threshold parameters. Other factors, such as total sample size and effect size of the general factor mean difference, seemed to have little influence on the estimation bias when DIF was present in threshold parameters. In comparison with the estimation bias of the general

factor mean difference in the binary data and the 3-point scale data in which all the noninvariant threshold parameters were constrained to be equal across groups for identification purposes, the estimation bias was smaller when two of the noninvariant threshold parameters were freely estimated for the 5-point scale data. When all noninvariant threshold parameters were constrained to be equal for the 5-point scale data, the estimation bias for the general factor mean difference became similar to those for the binary data and the 3-point scale data.

The relative estimation bias of the general factor mean difference for the conditions with DIF in threshold parameters is shown in Figure 10. As indicated in Figure 10, for the binary data and the 3-point scale data, the relative estimation bias of the general factor mean difference was nearly 20%, 30-40%, and 50-60% when the effect size was -0.1 and the DIF in the threshold parameters was 0.05, 0.10, and 0.15, respectively. Figure 10 showed that relative estimation bias decreased somewhat when two of the noninvariant threshold parameters for each item with DIF were freely estimated in the 5-point scale data, and the magnitude of decrease depended on the magnitude of DIF. When all threshold parameters with DIF were constrained to be equal for the 5-point scale data, the relative estimation bias of the general factor mean difference became similar to those for the binary data and the 3-point scale data. When the effect size was -0.2, the relative estimation bias became around half of those for the conditions with the effect size of -0.1.

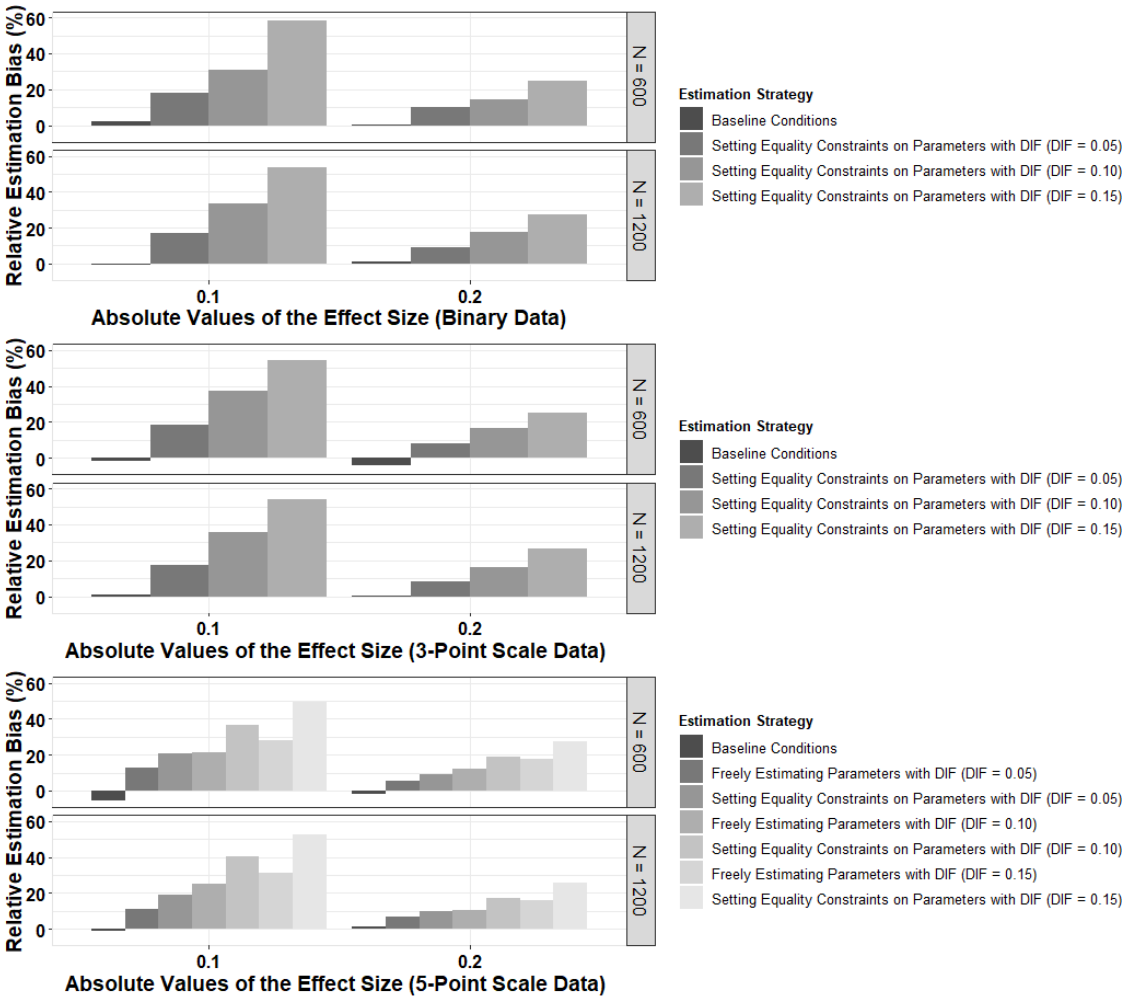

Figure 10 Relative Estimation Bias of the General Factor Mean in the Conditions with DIF in Threshold Parameters

Factors Influencing the Type I Error Rate/Power

The Type I error rate or power to detect the general factor mean difference for the conditions with DIF in general factor loadings are shown in Tables B13-B18 and Figure 11.

As shown in Tables B13 and B14, Type I error rates fell in the limits of .025 to .075 for all conditions with DIF in general factor loadings. In the conditions with DIF in general factor loadings, the Type I error rates were relatively lower for the 5-point scale data

(ranged from .028 to .043) and relatively higher for the 3-point scale data when the total sample size was 600 (ranged from .060 to .075). Whether or not the general factor loadings with DIF constrained to be equal seemed to have no influence on the Type I error rates.

Figure 11 indicated that the main factors influenced the power to detect the general factor mean difference in the conditions with DIF in general factor loadings were the total sample size and the effect size. As shown in Tables B13-B18, when the effect size was -0.1, empirical powers ranged from .168 to .222 and from .294 to .381 for conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .556 to .637 and from .822 to .882 for the conditions with total sample size of 600 and 1200, respectively. In addition to the total sample size and the effect size, the number of categories per item might also influence the values for the power to detect the general factor mean difference when DIF was present in general factor loadings. Specifically, for a given total sample size and effect size, powers were largest for the 3-point scale data and smallest for the 5-point scale data in most cases.

As indicated in Figure 5 and 11, in comparison with the correctly specified model, constraining the general factor loadings with DIF to be equal across groups produced more estimation bias in the general factor mean difference, but it had little influence on the power to detect the general factor mean difference.

Type I error rates regarding the general factor mean difference estimation for the conditions with specific factor loadings having DIF are shown in Tables B19 and B20. Type I error rates fell in the limits of .025 to .075 for all conditions with DIF in specific factor loadings. Similar to the conditions with DIF in general factor loadings, Type I error rates were relatively smaller for the 5-point scale data than for the 3-point scale data when

the DIF was present in specific factor loadings. Also, constraining the noninvariant specific factor loadings to be equal seemed to have no obvious impact on the Type I error rates.

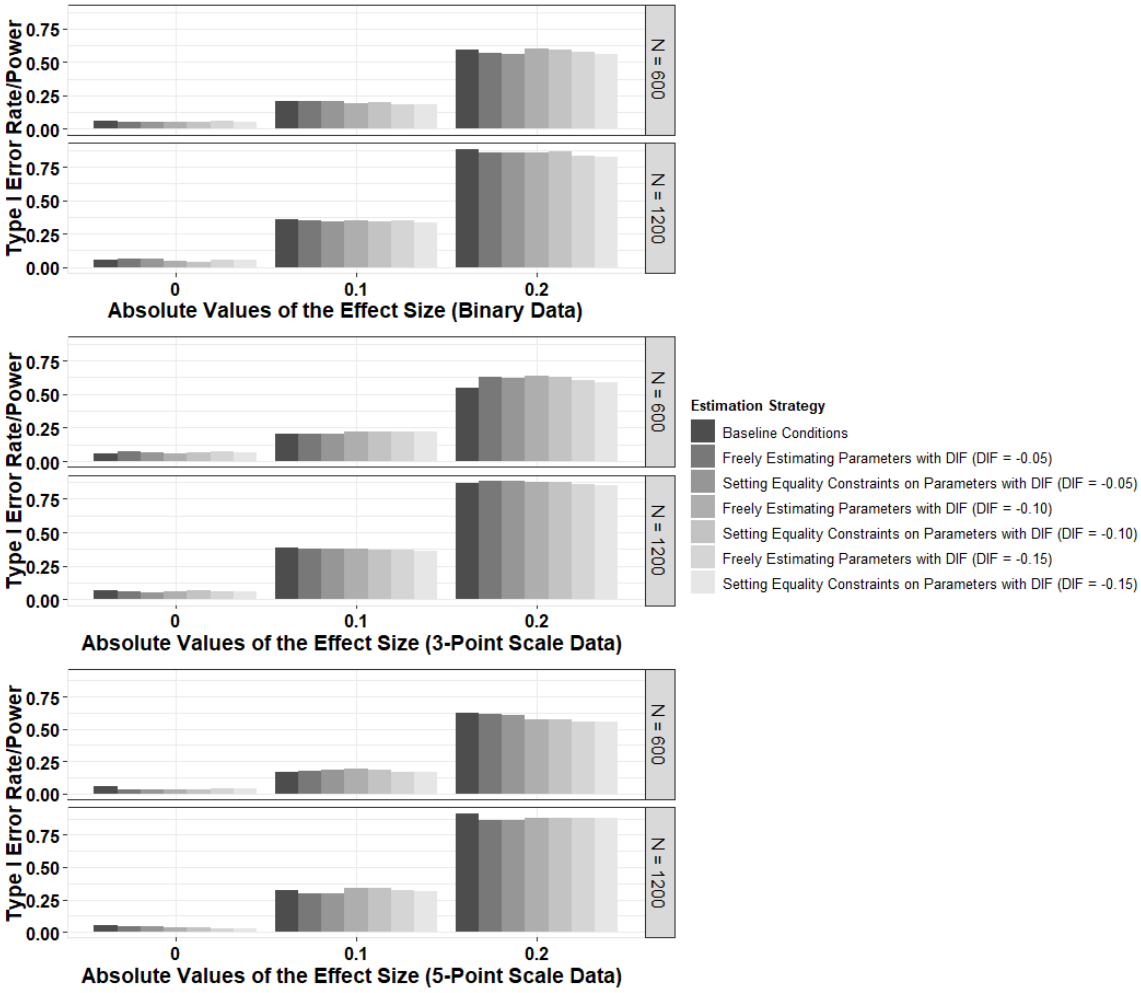

Figure 11 Type I Error Rate/Power to Detect the General Factor Mean Difference in the Conditions with DIF in General Factor Loadings

As shown in Figure 12, the dominant factors influencing the power to detect the general factor mean difference for the conditions with DIF in specific factor loadings were total sample size and effect size. When the effect size of the general factor mean difference was -0.1, the values of power ranged from .171 to .227 and from .314 to .400 for the

conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .571 to .642 and from .857 to .909 for the conditions with total sample size of 600 and 1200, respectively. For a given total sample size and effect size, in comparison with the conditions with DIF in general factor loadings, powers to detect the general factor mean difference fell in similar ranges for the conditions with DIF in specific factor loadings. The number of categories per item also slightly influence the power to detect the general factor mean difference when the DIF was present in specific factor loadings. For example, power seemed to be smaller for 5-point scale data than 3-point scale data in most cases. Also, whether the specific factor loadings with DIF were constrained to be equal across groups seemed to have little influence on the power to detect the general factor mean difference.

Type I error rates regarding the general factor mean difference estimation for the conditions with DIF present in threshold parameters are shown in Tables B25 and B26. When the total sample size was 600, the Type I error rates fell in the limits of .025 to .075 for the conditions with DIF in the threshold parameters of 0.05 and 0.10, and in the 5-point scale data with two of the noninvariant threshold parameters freely estimated, the Type I error rate also fell in the limits of .025 and .075 for the condition with the DIF of 0.15. When the total sample size was 1200, the Type I error rates fell in the limits of .025 to .075 in the 3-point scale data for the condition with DIF in the threshold parameters of 0.05 and in the 5-point scale data for the conditions with DIF in the threshold parameters of 0.05 and 0.10. Also, the Type I error rates fell in the limits of .025 to .075 seemed to be smaller for the 5-point scale data in comparison with those for the binary data and the 3-point scale data. All other Type I error rates for the conditions with DIF in threshold parameters were

inflated using .075 as the upper limit, and the magnitude of inflation depended on the magnitude of the DIF in the threshold parameters and the total sample size. Larger magnitude of DIF in the threshold parameters and larger total sample size were associated with more serious inflation of the Type I error rates regarding the latent mean difference estimation.

When Type I error rates fell beyond the limits of .025 to .075, the corresponding power to detect the general factor mean difference cannot be appropriately interpreted, so only the subset of the empirical detection rates for the nonzero effect size can be interpreted as power when DIF was present in threshold parameters. As shown in Tables B27-B30, the powers were mainly determined by the total sample size and the effect size. When the effect size of the general factor mean difference was -0.1, powers ranged from .234 to .331 and from .395 to .587 for the conditions with total sample size of 600 and 1200, respectively; when the effect size was -0.2, they ranged from .630 to .765 and from .921 to .963 for the conditions with the total sample size of 600 and 1200, respectively. In comparison with the powers for the conditions with DIF in general factor loadings or specific factor loadings, the powers for conditions with DIF in threshold parameters were obviously larger for a given total sample size and effect size. In addition to the total sample size and the effect size, the magnitude of DIF in the threshold parameters and the number of categories per item also influenced the power to detect the general factor mean difference. To be specific, larger magnitudes of DIF were associated with greater power, and the power was relatively smaller for the 5-point scale data than that for the 3-point scale data in general.

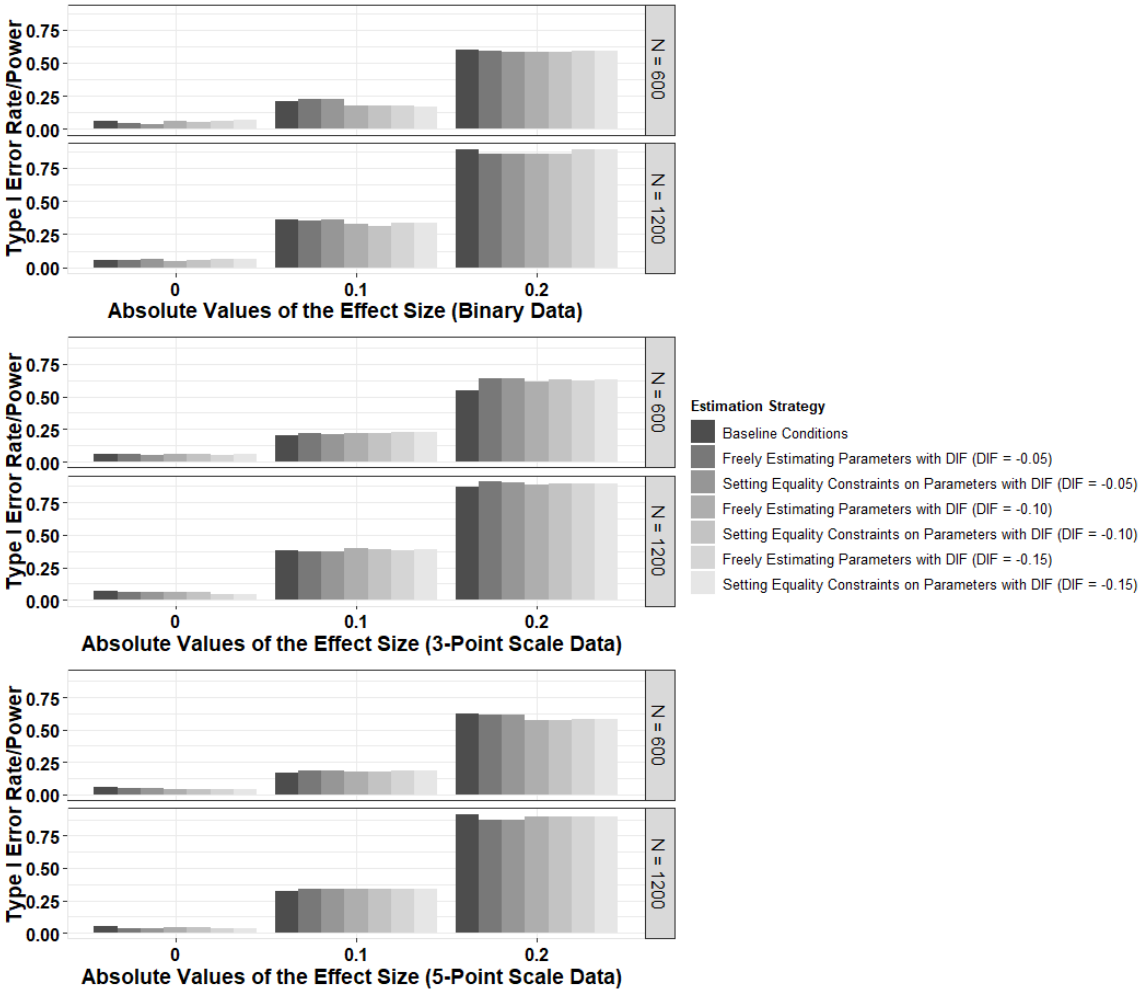

Figure 12 Type I Error Rate/Power to Detect the General Factor Mean Difference in the Conditions with DIF in Specific Factor Loadings

Estimated Variance

Similar to the No DIF conditions, in the conditions with DIF, the main factor influencing the estimated variance of the general factor mean difference was the total sample size (shown in Tables B13-B30). When the total sample size was 600, the estimated variances ranged from .007 to .009; when the total sample size was 1200, they ranged from .003 to .005. Estimated variances for these conditions with DIF fell in similar ranges to those for the No DIF conditions for a given sample size. Additionally, in comparison

with the corresponding correctly specified model, setting equality constraints on general factor loadings with DIF led to a smaller estimated variance in some conditions with the total sample size of 600 and only a few conditions with the total sample size of 1200, and the difference in the estimated variance for a given generated dataset was .001. When DIF was present in specific factor loadings or threshold parameters, estimated variances for the general factor mean difference remained the same after setting (more) equality constraints on the noninvariant parameters in comparison with the corresponding model with fewer equality constraints. Also, the estimated variance for the general factor mean difference seemed to be smaller for the data with more categories per item in general for a given total sample size in some of the cases. Given the level of precision reported (rounding to the thousandths place), not all differences in estimated variances were reported.

Coverage Rates of 95% Confidence Interval

In conditions with DIF in general factor loadings or specific factor loadings, almost all the coverage rates of the 95% confidence interval were above .950 no matter whether the parameters with DIF were freely estimated or not (shown in Tables B13-B24). When DIF was present in threshold parameters, most of the coverage rates of the 95% confidence interval were above .950 when the DIF = 0.05. When the DIF in the threshold parameters was 0.10 and 0.15, some coverage rates of the 95% confidence interval fell below .950 due to the serious estimation bias, and larger total sample size usually made them drop down more seriously.

Goodness of Fit Indices

The means of the goodness of fit indices (i.e., CFI, WRMR, and RMSEA) for the conditions with DIF were shown in Tables B31-B48.

Means of CFIs. As shown in Tables B31-B42, for the correctly specified models with all item parameters with DIF (i.e., noninvariant general factor loadings or noninvariant specific factor loadings) freely estimated, the means of CFIs were .999 with only one exception of .998 for a condition with the total sample size of 600. The results regarding the means of CFIs for these correctly specified models were similar to the baseline conditions in which there was no DIF in item parameters and all the item parameters were constrained to be equal across groups in analysis. After imposing equality constraints on the general factor mean difference for these correctly specified models, the means of CFIs did not change or even increased by .001 when the effect size of the general factor mean difference was zero. When the effect size was -0.1, the drop of the means in CFIs resulting from setting equality constraints on the general factor mean difference was .001 for almost all the correctly specified models. Only for the conditions involving binary data in which DIF was present in specific factor loadings and the total sample size was 1200, the means of CFIs did not change when the effect size was -0.1. When the effect size of the general factor mean difference was -0.2, the drop of the means in CFIs resulting from setting equality constraints on the general factor mean difference was .003 for binary data and .004 to .005 for polytomous data.

When DIF was present in general factor loadings, in comparison with the corresponding correctly specified model, constraining the general factor loadings with DIF of -0.05 to be equal across groups did not change the means of CFIs in most cases, and for two conditions involving binary data, they decreased by .001. When the DIF in the general factor loadings was -0.10, the drops of the means of CFIs due to setting equality constraints on the general factor loadings with DIF was .002, 0, and .001 for the binary, the 3-point

scale, and the 5-point scale data, respectively; when the DIF = -0.15, they decreased by .004, .001, and .002 for the binary, the 3-point scale, and 5-point scale data, respectively. In the models with general factor loadings having DIF constrained to be equal, imposing equality constraints on the general factor mean difference, the means of CFIs did not change or even increased when the effect size was zero; they decreased by .001 for some of the conditions with the effect size of -0.1, and they did not change or even increased for the rest; when the effect size was -0.2, the drops of CFIs were mostly .003 (with one of .002 and one of .004) for binary data and mostly .004 (with a few exceptions of .003) for polytomous data.

When DIF was present in specific factor loadings, constraining the noninvariant specific factor loadings to be equal only decreased the means of CFIs by .001 for most of the conditions with DIF of -0.15 and very a few conditions with DIF of -0.10. For the rest of the conditions (i.e., all conditions with DIF of -0.05, most conditions with DIF of -0.10, and a very few conditions with DIF = -0.15), the means of CFIs did not change due to setting equality constraints on the specific factor loadings with DIF. For the conditions with specific factor loadings having DIF constrained to be equal, after imposing equality constraints on the general factor mean difference, the means of CFIs did not change or even increased when the effect size was zero; they did not change or decreased by .001 when the effect size was -0.1; and when the effect size was -0.2, the drop of the means of CFIs was .003 or .004 for binary data and mostly .004 (with one exception of .005) for polytomous data.

Results regarding the means of CFIs for the conditions with DIF in threshold parameters are shown in Tables B43-B48. When the DIF was present in threshold

parameters, none of the models could be correctly specified in the current study because equality constraints had to be placed on the only threshold parameter for each item in binary datasets and at least two threshold parameters for each item in polytomous datasets for identification purposes. As shown in Tables B43-B48, setting equality constraints on the only threshold parameter for the binary data or at least two threshold parameters for the polytomous data, the means of CFIs were .999, .999 with some exceptions of .998, and mostly .998 with a few exceptions of .999 when the DIF in the threshold parameters was 0.05, 0.10, and 0.15, respectively. For the 5-point scale data, after constraining more noninvariant threshold parameters to be equal, the means of CFIs did not change when the DIF = 0.05, and they decreased by .001 for most of the conditions with DIF of 0.10 and 0.15. In the models with DIF in threshold parameters, constraining the general factor mean difference to be zero, the means of CFIs did not change or even increased when the effect size of the general factor mean difference was zero; they decreased by .001 or .002 for conditions with the effect size of -0.1 and by .004-.007 for conditions with the effect size of -0.2.

Means of WRMRs. As shown in Tables B31-B42, when the DIF was present in the general factor loadings or specific factor loadings, for the correctly specified models with all item parameters with DIF freely estimated, the means of WRMRs ranged from .938 to .972 and from .786 to .833 for binary and polytomous data, respectively. Additionally, the means of WRMRs also increased slightly as the total sample size got larger. After imposing equality constraints on the general factor mean difference of zero, the means of WRMRs ranged from .969 to 1.001 and from .837 to .896 for binary data and polytomous data, respectively. When the effect size was -0.1, the means of WRMRs for the constrained

model ranged from 1.003 to 1.065 and from .892 to 1.053 for binary data and polytomous data, respectively; when the effect size was -0.2, they ranged from 1.094 to 1.246 and from 1.040 to 1.442 for binary data and polytomous data, respectively. The increase in the means of WRMRs due to setting equality constraints on the general factor mean difference was relatively most substantial for the 5-point scale data.

When the DIF was present in general factor loadings, after constraining the general factor loadings with DIF of -0.05 to be equal, the means of WRMRs ranged from .995 to 1.029 and from .812 to .857 for binary data and polytomous data, respectively; when the DIF = -0.10, they ranged from 1.048 to 1.120 and from .824 to .898 for binary data and polytomous data, respectively; when the DIF = -0.15, they ranged from 1.124 to 1.257 and from .853 to .967 for binary and polytomous data, respectively.

When DIF was present in specific factor loadings, after constraining the specific factor loadings with DIF of -0.05 to be equal, the means of WRMRs ranged from .984 to .994 and from .808 to .848 for binary and polytomous data, respectively; when the DIF = -0.10, they ranged from .992 to 1.015 and from .817 to .866 for binary data and polytomous data, respectively; when the DIF = -0.15, they ranged from 1.009 to 1.046 and from .836 to .905 for binary data and polytomous data, respectively.

As shown in Tables B31-B42, for a given effect size of the general factor mean difference, the increase in the means of WRMRs resulting from imposing equality constraints on the general factor mean difference in the models with noninvariant general factor loadings or noninvariant specific factor loadings constrained to be equal across groups was similar to those in the corresponding correctly specified models.

When DIF of 0.05 was present in threshold parameters, setting equality constraints on the only threshold parameter of each item for the binary data or at least two threshold parameters of each item for the polytomous data, the means of WRMRs ranged from .980 to .993 and from .785 to .843 for binary and polytomous data, respectively; when the DIF = 0.10, they ranged from .990 to 1.014 and from .798 to .878 for binary data and polytomous data, respectively; when the DIF = 0.15, they ranged from 1.007 to 1.048 and from .825 to .935 for binary data and polytomous data, respectively (shown in Tables B43-B48). For a given magnitude of DIF in threshold parameters, if only constraining two threshold parameters of each item to be equal across groups for identification purpose, the means of WRMRs were smaller for the 5-point scale data than those for the 3-point scale data. For the 5-point scale data with DIF in threshold parameters, after constraining all threshold parameters to be equal, the means of WRMRs ranged from .817 to .860, from .844 to .910, and from .889 to .993 when the DIF = 0.05, 0.10, and 0.15, respectively. As shown in Tables B43-B48, for a given effect size of the general factor mean difference, there were not much differences of the increase in the means of WRMRs resulting from imposing equality constraints on the general factor mean difference between the conditions with DIF in threshold parameters and the conditions with DIF in general factor loadings or specific factor loadings.

Means of RMSEAs. As shown in Tables B31-B42, when DIF was present in the general factor loadings or specific factor loadings, for the correctly specified models with all the item parameters with DIF freely estimated, the means of RMSEAs ranged from .007 to .015, with their values slightly influenced by the total sample size and number of categories per item. Smaller sample sizes and larger number of categories per item were

associated with larger values of RMSEAs. When the effect size of the general factor mean difference was zero, the means of RMSEAs decreased after imposing equality constraints on the general factor mean difference in most of the cases, with very few exceptions in which they did not change. When the effect size of the general factor mean difference was nonzero, the increase in the means of RMSEAs resulting from imposing equality constraints on the general factor mean difference ranged from .001 to .006 and from .013 to .023 for the conditions with effect size of -0.1 and -0.2, respectively, which was slightly influenced by the total sample size and the number of categories per item.

When DIF was present in the general factor loadings, constraining noninvariant general factor loadings to be equal, the means of RMSEAs increased by 0 to .003, .002 to .011, and .005 to .020 for the conditions with the DIF of -0.05, -0.10, and -0.15, respectively. The magnitude of the increase in the means of RMSEAs was influenced by the total sample size and the number of categories per item. After imposing equality constraints on the general factor loadings with DIF, the means of RMSEAs increased most substantially for the binary data. Also, larger sample sizes were associated with larger increases in the means of RMSEAs resulting from constraining the noninvariant general factor loadings to be equal. When setting equality constraints on the nonzero general factor mean difference, the increases in the means of RMSEAs for the conditions with noninvariant general factor loadings constrained to be equal were smaller than those for the corresponding conditions with correctly specified models in general.

When the DIF was present in specific factor loadings, after setting equality constraints on these noninvariant specific factor loadings, the increase in the means of RMSEAs ranged from 0 to .001, from .001 to .004 and from .003 to .007 for the conditions

with DIF of -0.05, -0.10, and -0.15, respectively. Also, after constraining the nonzero general factor mean difference to be zero, the increase in the means of RMSEAs for the conditions with equality constraints on the specific factor loadings with DIF was slightly smaller than those for the corresponding conditions with correctly specified models in most of the cases.

When the DIF was present in threshold parameters, setting equality constraints on the only threshold parameter of each item for the binary data or at least two threshold parameters of each item for the polytomous data, the means of RMSEAs ranged from .010 to .014, from .011 to .017, and from .013 to .021 for the conditions with DIF of 0.05, 0.10, and 0.15, respectively. After constraining more noninvariant threshold parameters to be equal for the 5-point scale data, the means of RMSEAs increased by .001, .003 (with one exception of .002), and .005 (with one exception of .004) when DIF = 0.05, 0.10, and 0.15, respectively. When the effect size of the general factor mean difference was nonzero, the increase in the means of RMSEAs resulting from imposing equality constraints on the general factor mean difference ranged from .002 to .009 and from .016 to .024 for the conditions with effect sizes of -0.1 and -0.2, respectively.

Chapter 4: Discussion

Overview

In educational, psychological, and social science disciplines, bifactor models are increasingly applied because they often serve as the most appropriate representations for measurement systems in which relatively broader constructs (e.g., depression) additionally may have multiple, narrower facets (e.g., negative mood, social withdrawal, poor cognitive functioning, etc.) that should be modeled. Similarly, cognitive tests for a general domain such as reading comprehension may include multiple testlets, which are clusters of items based on common stimuli (e.g., reading passages) or text type (e.g., narrative vs. expository) that create additional dimensionality in the data. Despite the prevalence of bifactor data, only a few methodological studies focusing on multiple-group bifactor models have been undertaken (e.g., Fukuhara & Kamata, 2011; Jeon et al., 2011; Cai et al., 2011), and these studies focused on the DIF detection or item parameter recovery. Given that (latent) mean differences between populations are often of interest to researchers from different disciplines, and that for bifactor data, researchers are often interested in population differences in the distributions of the primary trait, the current simulation study examined the performance of several approaches to estimating the latent mean difference of the general factor for ordinal, bifactor data.

The approaches involved in the current study varied in terms of the choice of analysis models (unidimensional models vs. bifactor models), estimators (the WLSMV estimator or the MLR estimator), and whether equality constraints were imposed on the item parameters with DIF.

Results showed that bias in the general factor mean difference estimation was produced mainly when the model was misspecified by fitting the generated bifactor data

using unidimensional models or setting equality constraints on item parameters with DIF. Treating ordered categorical data as continuous did not yield estimation bias in the general factor mean difference. Although the estimation bias of the general factor mean difference was influenced by different analysis models to varying degrees, the most dominant factors that influenced Type I error rates or powers to detect the general factor mean difference were total sample size and effect size. As expected, the more complicated models usually produced less estimation bias but they also had less estimation precision, demonstrating the tradeoff between the estimation bias and estimated variance.

Robustness of Latent Mean Difference Estimation under Unidimensional IRT Models to Multidimensional Violation

As stated by Reise et al. (2010), although most IRT models applied today are unidimensional models, strict unidimensional models rarely exist, and researchers are usually more interested in whether the data are sufficiently unidimensional to satisfy a weak form of local independence assumption. On one hand, in addition to prevalent applications, unidimensional models might be preferred due to their theoretical simplicity. On the other hand, researchers want to avoid the problems due to the violation of the unidimensionality assumption. Given that there are no absolute and consistent criterions to determine whether a dataset is sufficiently unidimensional, in single-group IRT practice, researchers are usually more concerned about the impact on item parameter estimates that may result from fitting potentially multidimensional data using a unidimensional model. Reise et al. (2010) proposed comparing factor loadings of a unidimensional model with the general factor loadings of the corresponding bifactor model to figure out whether there are problems in item parameter estimates due to violation of unidimensionality assumption.

DeMars (2006) also compared item parameters obtained from bifactor models and unidimensional models for generated bifactor datasets. Results from these studies indicated the general factor loadings (corresponding to general discrimination parameters in IRT models) were distorted after fitting unidimensional models to bifactor data. Also, DeMars (2006) pointed out the recovery of difficulty parameters did not appear to be influenced by fitting the generated bifactor data using a unidimensional model.

Similarly, in multiple-group IRT models, instead of discussing the strength or weakness of the criteria to determine the degree of unidmensionality, it is more important to understand the consequence on DIF detection or subsequent analysis such as latent mean comparison of the general factor resulting from fitting unidimensional models to the generated bifactor datasets. Fukuhara and Kamata (2011) conducted a study in which multiple-group bifactor data were generated with DIF in item difficulty parameters and analyzed using both bifactor models and unidimensional models. They found DIF could be better detected using bifactor models in comparison with the corresponding unidimensional model. In the current study, the consequence on the general factor mean difference estimation resulting from the violation of unidimensional assumption was the focus. Results showed the estimation bias for the correctly specified models (i.e., fitting the bifactor data using bifactor models) was around 0 in general, and that positive estimation bias was produced when fitting unidimensional models to the generated bifactor datasets. The magnitude of the increase in estimation bias of the general factor mean difference resulting from the violation of unidimensional assumption mainly depended on the effect size of the general factor mean difference and the degree of unidimensionality (i.e., the sizes of specific factor loadings) of the generated data in terms of ECV; it was also slightly

influenced by the selection of the estimator (i.e., MLR vs. WLSMV). More specifically, when the effect size of the general factor mean difference was 0, there was no obvious change in the estimation bias due to fitting the bifactor data using a unidimensional model; when the effect size was 0.2, the increase in the estimation bias was almost two times that for the conditions with the effect size of 0.1. When generating the data, the ECV was around .84 and .66 for the data with high degree and low degree of unidimensionality (i.e., less multidimensional and more multidimensional data), respectively, suggesting that 84% and 66% of the explained common variance in the data was attributed to the general factor. If fitting a unidimensional model to the generated bifactor data with nonzero effect size in the general factor mean difference, the absolute values of the relative bias in the general factor mean difference were around 2.5-5.0% for less multidimensional data, and these values usually reached 10-15% for more multidimensional data. The relative estimation bias did not appear to be influenced by the effect size of the latent mean difference.

For the limited-information estimation method (i.e., the WLSMV estimator) applied in the current study, the threshold parameters and the tetrachoric or polychoric correlations were estimated first either simultaneously or separately, and then a CFA model was fitted to the tetrachoric or polychoric correlations (Rhemtulla, Brosseau-Liard, & Savalei, 2012; Wirth & Edwards, 2007), so the change of the model structure (unidimensional model vs. bifactor model) would probably influence the model-implied tetrachoric or polychoric correlations by using a different set of parameter estimates. As mentioned previously, fitting unidimensional models to bifactor data usually made the factor loadings (corresponding to discrimination parameters) distorted, so it can be inferred that the function of the estimation bias in the general factor mean difference influenced by

the violation the unidimensional assumption was mainly through the distorted factor loadings. Note that in the current study, even for conditions with low degree of unidimensionality, all the specific factor loadings (i.e., 0.5) were smaller than the general factor loadings (i.e., 0.7). Thus, for the cases with relatively stronger specific factors, the distortion of the factor loadings likely would be even more serious, which would result in more estimation bias in the general factor mean difference.

In addition to estimation bias, the mean squared error (i.e., MSE) or the root mean square error (RMSE), the combination of the estimation accuracy and estimation precision, was also of great importance in evaluating an estimation procedure. As shown in the results, the main factor influencing estimated variances of the general factor mean difference was sample size. Although the increase in estimation bias of the general factor mean difference resulting from fitting the bifactor datasets using unidimensional models was not influenced obviously by the sample size, the estimation precision got worse for the conditions with smaller total sample sizes. Also, the results of the current study indicated that analysis with unidimensional models usually led to estimated variances for the general factor mean difference that were smaller by .001 or .002 in comparison with the corresponding bifactor models for the same generated dataset. These reductions in the estimated variance were much larger than the respective increases in the squared estimated bias (i.e., around .0006 at maximum) due to analyzing bifactor data with unidimensional models in the research conditions of the current study Thus, in the current study, the consequence on the estimation for the general factor mean difference resulting from fitting unidimensional models to the generated bifactor model might be acceptable in terms of MSE, which differs from DeMars (2006)'s results regarding the RMSE in item parameter estimates. The

difference of the conclusions between the current study and DeMars' study (2006) suggests that whether the consequence of fitting bifactor data with a unidimensional model was acceptable depends on the focal estimated parameters. In addition, the magnitude of the manipulated factors might also influence these conclusions. Given the respective factors influencing the estimation accuracy and estimation precision of the general factor mean difference estimation discussed above, it can be inferred that the estimation bias resulting from fitting the unidimensional model to a potentially bifactor dataset might become the more dominant factor in determining the MSE or RMSE for the general factor mean difference estimation as the ECV decreases and the effect size increases.

Estimation with Robust Maximum Likelihood vs. Categorical Variable Methodology for the Ordinal Bifactor Data

In practice, researchers often treat ordinal data as continuous for the following two reasons: first, some researchers are more familiar with estimation methods for continuous data; second, the numerical coding of the ordinal data in an ascending order makes them look similar to continuous data (Rhemtulla et al., 2012). However, ignoring the non-continuity and non-normality of the ordinal data might contribute to estimation problems. Whether ordinal data can be treated as continuous has been explored by many researchers (e.g., Rhemtulla et al., 2012; Stark et al., 2006), and there is some agreement that continuous data estimation strategies perform as well as categorical data estimation strategies if the number of categories is large enough (i.e., 5 or more). Researchers also explored other factors that influenced the choice between the continuous variable methodology and categorical variable methodology for ordered-categorical data. For example, based on the simulation study of Rhemtulla et al. (2012), robust maximum

likelihood estimation might not be appropriate for the ordinal data with asymmetric threshold parameters. Also, researchers (e.g., Rhemtulla et al., 2012; Stark et al., 2006) recommended treating ordinal data with 5 or more categories as continuous when the sample size is small (e.g., 150 for single-group analysis or 500 for multiple-group analysis).

For multiple-group ordinal data, several studies were conducted to compare continuous and categorical estimation strategies in terms of DIF detection (e.g., Desa, 2014; Flowers et al., 2002; Meade & Lautenschlager, 2004; Stark et al., 2006). There were no consistent conclusions regarding the performance of the continuous approach in the DIF detection from these simulation studies. Some researchers (e.g., Desa, 2014; Meade & Lautenschlager, 2004) pointed out that the continuous approach using ML or MLR estimation was unable to correctly detect the DIF in threshold parameters because there were not exact corresponding parameters in continuous CFA models for the threshold parameters, while some other researchers (Stark et al., 2006) showed that continuous CFA model with the ML estimator performed similarly in detecting DIF in both loading parameters and threshold parameters as the IRT model.

In the current simulation study, however, DIF detection was not the focus. Regarding the estimation bias of the general factor mean difference, both the categorical approach with the WLSMV estimator and the continuous approach with the MLR estimator performed acceptably as long as the model was correctly specified using bifactor models. When the generated more multidimensional data (i.e., specific factor loadings = 0.5) was fitted with unidimensional models, large estimation bias in the general factor mean difference was produced for both the MLR estimator and the WLSMV estimators, but the

estimation bias was smaller when implementing the MLR estimator than that when implementing the WLSMV estimator.

As indicated in the results, the Type I error rate and power to detect the general factor mean difference were slightly influenced by the selection of between the Satorra-Bentler scaled chi-square difference test for MLR estimator and the DIFFTEST for the WLSMV estimator. More specially, for the 3-point scale data and the 5-point scale data, the Type I error rate or the power obtained through the Satorra-Bentler scaled chi-square difference test for the MLR estimator seemed to be a little smaller than that obtained through the DIFFTEST for the WLSMV estimator in general.

In summary, in order to obtain higher power to detect the general factor mean difference, the WLSMV estimator was recommended over the MLR estimator although the improvement was limited. The choice between the MLR estimator and the WLSMV estimator had no substantial influence on the estimation accuracy of the general factor mean difference except in the severe misspecification conditions. Inconsistent with our expectations, total sample size (i.e., 600 or 1200) and number of categories per item (i.e., 3 or 5) seemed to have no influence in the preference between the MLR estimator and the WLSMV estimator.

Goodness of Fit Indices for the No DIF Conditions

Previous studies (e.g., Chen, 2007; Cheung & Rensvold, 2002) have shown that the change of goodness of fit indices can be applied for testing different levels of measurement invariance. In the current study, the means of CFI, WRMR/SRMR and RMSEA were reported for each of the analysis models, but changes in goodness of fit indices per se were not evaluated. The results revealed that all the goodness of fit indices (i.e., CFI,

WRMR/SRMR, RMSEA) suggested nearly perfect fit when bifactor models fit to the generated bifactor data except that the means of WRMRs were a little bit larger than 1 for some conditions involving binary data. Further, the selection of the estimator (i.e., the MLR estimator or the WLSMV estimator) had little influence on the model fit reflected by these indices if the model was correctly specified using bifactor models.

Imposing equality constraint on the general factor mean difference of zero, the means of CFIs or RMSEAs did not change or suggested a better model fit, while the means of SRMRs and WRMRs increased a little bit. When the effect size was nonzero, all the means of these goodness of fit indices indicated a poorer model fit after constraining the general factor mean difference to be zero in general, but the degree of changes varied for these indices.

To be specific, the reductions in means of CFIs due to constraining the nonzero general factor mean difference to be zero were the minimal (i.e., less than .01 even when the effect size of the general factor mean difference was -0.2), and they were especially small when the MLR estimator was applied (i.e., .001 when the effect size was -0.2 and 0 in most of the cases when the effect size was -0.1), so CFIs might not be sensitive for determining the significance of the nonzero effect size for the general factor mean difference in bifactor models.

The WRMRs seemed to be most sensitive to the nonzero effect size of the general factor mean difference when it was constrained to be zero in terms of both the changes of the mean values (i.e., around 0.1-0.2 and 0.2-0.6 for the effect size of -0.1 and -0.2) and the absolute mean values (i.e., larger than 0.9 when the effect size was -0.1 and larger than 1 when the effect size was -0.2) for polytomous data, but it should be noted that the means

of WRMRs also increased a little bit when constraining to zero the latent mean difference with effect size of zero. Another weakness of the WRMRs was that both their changes in means and their absolute values in means were influenced by the data generation conditions such as the sample size, the number of categories per item, and even the degree of unidimensionality.

The increase in the means of SRMRs resulting from constraining the general factor mean difference of -0.1 to be zero was just a little bit larger than that due to constraining to zero the general factor mean difference with effect size of zero. When constraining the general factor mean difference of -0.2 to be zero, the increase in the means of SRMRs became obvious (i.e., around 0.01), but the absolute values for the means of SRMRs still suggested very good fit.

The changes in the means of RMSEAs were somewhat sensitive to the nonzero effect size of the general factor mean difference when the WLSMV estimator was applied (i.e., .002-.006 and .013-.025 for effect size of -0.1 and -0.2, respectively), and they did not provide much information in detecting the nonzero effect size when the MLR estimator was applied (0-.001 and .002-.005 for effect size of -0.1 and -0.2). Thus, when the WLSMV estimator was applied, the goodness of fit indices (i.e., CFI, WRMR, RMSEA) tended to be more sensitive in detecting the nonzero effect size of the general factor mean difference for the bifactor models in comparison with those for the conditions with the MLR estimator used (i.e., CFI, SRMR, RMSEA).

In the No DIF conditions, all the goodness of fit indices (i.e., CFI, WRMR/SRMR, RMSEA) were very sensitive to model misspecification of fitting the unidimensional models to the generated bifactor data, and the decreases in the means of CFIs were

especially substantial when the MLR estimator was applied. However, CFIs and RMSEAs cannot provide any help in detecting the nonzero effect size of the general factor mean difference in unidimensional models because their means did not change obviously or even suggested a better fit after imposing equality constraints between groups on the general factor mean difference of -0.1 or -0.2. Thus, researchers should use caution when using the change of goodness of fit indices (i.e., CFI or RMSEA) for comparing model fits if the less constrained model was already misspecified.

The Impact of DIF on the General Factor Mean Difference Estimation in Bifactor Models

Although it is ideal to conduct latent mean comparisons based only on the invariant items, previous simulation studies have suggested that perfect recovery of the DIF was hard to achieve with commonly applied DIF detection methods, especially when the magnitude of the DIF or the sample size was not large enough (e.g., Narayanan & Swaminathan, 1996; Sweeney, 1996). Consistent with results from previous studies (e.g., Hancock et al., 2000; Yang, 2008), when the item parameters with DIF were freely estimated, the estimation accuracy of the latent mean difference was not adversely affected by the DIF; when failing to account for DIF, more estimation bias would be produced in comparison with the correctly specified model. Different from previous simulation studies (e.g., Beuckelaer & Swinnen, 2018) in which the impact of ignoring noninvariance on latent mean comparisons focused on models with simple structures, this study sought to evaluate estimation of the general factor mean difference within bifactor models with DIF in different item parameters. I found that the extent of impact of failing to account for DIF on estimation of

the general factor mean difference largely depended on the type of parameters generated to have DIF.

To be specific, as shown in the results, when the DIF was present in general factor loadings, in comparison with the No DIF conditions, there was no more estimation bias of the general factor mean difference produced if the general factor loadings with DIF were freely estimated. In comparison with this corresponding correctly specified model, when the general factor loadings with DIF were constrained to be equal in the analysis model, the increase in the estimation bias was positive and substantial for conditions with nonzero effect sizes, but no increase was observed when the effect size was zero. The increase in the estimation bias resulting from setting equality constraints on the general factor loadings with DIF for the conditions with effect size of -0.2 was about two times that for the conditions with effect size of -0.1; the increase in the estimation bias for the conditions with DIF magnitude of -0.10 and -0.15 was about 2 or 3 times that for the conditions with DIF magnitude of -0.05.

To calculate the relative estimation bias, the estimation bias was divided by the true parameter value, and therefore, the relative estimation bias was only influenced by the magnitude of DIF. For the misspecified conditions with noninvariant general factor loadings constrained to be equal across groups, the absolute values of the relative estimation bias for the general factor mean difference were around 5% when the magnitude of DIF was -0.15; they may be regarded as negligible for conditions with smaller DIF.

When the same degree of DIF was present in specific factor loadings, there was no more estimation bias produced than in the comparable No DIF conditions regardless of whether the specific factor loadings with DIF were freely estimated or not. This suggests

that setting equality constraints on the noninvariant specific factor loadings had little influence on the estimation bias of the general factor mean difference.

Also for DIF present in threshold parameters comparable to that present for factor loadings, the estimation bias of the general factor mean difference became very substantial even when the magnitude of DIF was 0.05. When generating the data, for the items with DIF in threshold parameters, a constant (i.e., 0.05, 0.10, or 0.15) was added to all threshold parameters. It worth noting that the only threshold parameter per item for the binary data and two of the threshold parameters per item for the polytomous data needed to be constrained to be equal between groups for identification purposes. Thus, in practice, even if DIF in threshold parameters was correctly detected, some threshold parameters with DIF had to be constrained to be equal to identify the model in multiple-group categorical CFA models. In the current study, all threshold parameters with DIF had to be constrained to be equal for the binary data and the 3-point scale data; for the 5-point scale data, half of the threshold parameters with DIF had to be constrained to be equal and the other half were either freely estimated or constrained to be equal depended on different research conditions.

The results showed that the magnitude of DIF was the main factor influencing the estimation bias of the general factor mean difference resulting from setting equality constraints on the threshold parameters with DIF. When the magnitude of DIF in the threshold parameters was 0.10 or 0.15, the estimation bias in the general factor mean difference was around 2 or 3 times that for the conditions with DIF magnitude of 0.05, respectively, and all the estimation bias was substantially negative. For a given magnitude of DIF in threshold parameters, the estimation bias of the general factor mean difference resulting from constraining the noninvariant threshold parameters to be equal was similar

across all the effect sizes (i.e., 0, -0.1, or -0.2) and datasets with different number of categories per item as long as all the noninvariant threshold parameters were constrained to be equal. For the 5-point scale data, when freely estimating half of the threshold parameters with DIF, there was an obvious decrease in the estimation bias in comparison with the conditions with all the noninvariant threshold parameters were constrained to be equal.

When the effect size was -0.1, the values of the relative estimation bias of the general factor mean difference resulting from setting equality constraints on the noninvariant threshold parameters were nearly 20%, 30-40%, and 50-60% for DIFs of 0.05, 0.10, and 0.15, respectively, and they reduced by half when the effect size was -0.2. In general, even a very small DIF in the threshold parameters made the estimation bias of the general factor mean difference substantial. To illustrate conceptually how small the difference in response frequencies might be for the DIF of 0.05 in threshold parameters, consider an example for the 3-point scale data. In the reference group, the threshold parameters were -0.5 and 0.5, suggesting that about 31%, 38%, and 31% of the normally distributed latent response variates (M = 0, SD = 1) fell in categories 1, 2, and 3, respectively. If the magnitude of DIF was 0.05, the threshold parameters for the noninvariant items in the focal group would be -0.45 and 0.55, suggesting that about 33%, 38%, and 29% of the normally distributed latent response variates (M = 0, SD = 1) fell in each of the three categories when the mean difference in latent response variates was not considered. If constraining the threshold parameters with DIF to be equal, the falsely estimated threshold parameters would influence the estimation of the mean for the latent

response variates in the focal group, which would further influence the estimation of the general factor mean difference.

In the current study, the DIF was manipulated in general factor loadings, specific factor loadings, and threshold parameters. If corresponding to the item parameters in GRMs using Equation 16, when the DIF of a negative value (i.e., -0.05, -0.10, or -0.15) was present in the general factor loading for an item, this item would have relatively lower ability to discriminate individuals' differences in the general factor for the focal group in comparison with the reference group; its specific factor discrimination parameter and item intercept also shifted accordingly. Similarly, when the DIF of a negative value (i.e., -0.05, -0.10, or -0.15) was present in the specific factor loading for an item, this item would have relatively lower ability to discriminate individuals' differences in the specific factor for the focal group in comparison with the reference group; its general factor discrimination parameter and item intercept also shifted accordingly. With respect to the DIF in threshold parameters, a positive value (i.e., 0.05, 0.10, or 0.15) was added to each of the threshold parameters for the chosen noninvariant items in the focal group, suggesting these items were relatively more difficult for the examinees in the focal group. The effect size of the general factor mean difference was manipulated as 0, -0.1, or -0.2. The negative values in the effect size suggested the focal group's overall performance in the general factor were worse than the reference group's performance. As shown above, the positive estimation bias for the general factor mean difference with nonzero effect size produced by setting equality constraints on the general factor loadings with DIF means that the absolute general factor mean difference was underestimated after constraining the noninvariant general factor loadings to be equal. The negative estimation bias for the general factor mean

difference resulting from setting equality constraints on the threshold parameters with DIF means that the absolute general factor mean difference was overestimated after constraining the noninvariant threshold parameters to be equal.

With respect to the Type I error rate or power to detect the general factor mean difference for bifactor models in the conditions with DIF, DIF in factor loadings did not influence the values of the Type I error rate or power no matter whether the noninvariant loadings were freely estimated or not. However, when DIF was present in threshold parameters, Type I error rates were inflated in many conditions and, accordingly, the corresponding power could not be appropriately interpreted. Even the powers for conditions in which their corresponding Type I error rates fell in the designated limits (i.e., .025-0.075) were still obviously larger than corresponding baseline conditions in which no DIF was simulated, and they were influenced by the magnitude of DIF. Thus, researchers should be cautious when using significance tests for the general factor mean difference in bifactor models when there was DIF in threshold parameters.

Goodness of Fit Indices for the Conditions with DIF

The results showed that the performance of the goodness of fit indices (i.e., CFI, WRMR, RMSEA) for the correctly specified models with noninvariant item parameters freely estimated was similar to that for the corresponding baseline conditions discussed earlier. None of these fit indices seemed to worsen obviously, on average, when incorrect equality constraints were imposed on the item parameters with DIF, which suggested there may be difficulty in detecting the DIF using goodness of fit indices and incremental changes in these indices for multiple-group, bifactor, ordered-categorical CFA models.

Among these indices, the CFIs were least sensitive to the model misspecification due to ignoring the DIF. The smallest mean CFI among all the conditions with DIF was .995, assuming the general factor mean difference was freely estimated, meaning the decrease in the mean CFI cannot exceed .005 regardless of the type of item parameters with DIF and the magnitude of DIF.

The increase in the means of WRMRs resulting from setting equality constraints on noninvariant item parameters seemed to be relatively obvious, but it should be noted that the means of WRMRs would typically increase if the model became more constrained, and that the changes in the means of WRMRs were influenced by data generation conditions unrelated to the magnitude of DIF. As mentioned previously, values of WRMRs less than 0.9 or 1.0 suggest a good fit (Yu & Muthén, 2002; Yu, 2002). The absolute values of the means of WRMRs for binary data fell between 0.9 and 1 for correctly specified models and they were larger than 1 in some of the conditions when the DIF was ignored; for polytomous data, they were usually less than 0.9 even when the noninvariant item parameters were constrained to be equal, and they only fell between 0.9 and 1for some of the conditions with large DIF (usually 0.15) ignored and the total sample size of 1200.

The means of RMSEAs usually increased as the noninvariant item parameters were constrained to be equal, however, the largest mean of RMSEAs was .028 among all the conditions with DIF if the general factor mean difference was freely estimated, suggesting the increase in the means of RMSEAs due to constraining noninvariant item parameters to be equal cannot be very large. Among all the conditions with DIF, the RMSEAs were relatively most sensitive to setting equality constraints on the general factor loadings with DIF for binary data.

In summary, although constraining item parameters with DIF to be equal across groups might produce a different degree of estimation bias for the general factor mean difference, the goodness of fit indices for each model usually suggested good model fit regardless of the magnitude of estimation bias. As pointed out by Reise (2012), it might not be appropriate to use goodness of fit indices for linear CFA models to evaluate non-linear IRT models because they are estimated based on different assumptions. So a reason that these goodness of fit indices applied in the current study were not very sensitive to noninvariance of item parameters for ordinal bifactor data might be that they were developed for linear models rather than non-linear models.

Limitations and Future Studies

First, when exploring the impact of fitting unidimensional models to bifactor data on the general factor mean difference estimation, the absolute value for the general factor mean difference was underestimated due to the model misspecification. According to analysis of the estimation procedure and results from previous simulation studies (e.g., DeMars, 2006; Reise et al., 2010), it was inferred that the estimation bias in the general factor mean difference was produced through the distorted factor loading estimates resulting from violation of the unidimensionality assumption. However, there were no consistent conclusions about the direction of the distortion of the factor loading estimates, so additional data generation conditions—such as the number of testlets, the number of items within each testlet, the correlations among the testlets, and the loadings on each testlet—might be informative for understanding how the factor loadings are distorted due to fitting unidimensional models to bifactor data.

Second, in the current study, the estimation of the general factor mean difference was robust to fitting unidimensional models to bifactor data in terms of MSE (i.e., sum of the squared estimation bias and estimated variance). However, as revealed in the results, larger effect size of the general factor and lower degree of unidimensionality would result in larger estimation bias, while estimated variance was usually determined by the sample size and the degrees of freedom, so the estimation bias still has the potential to increase while holding the estimated variance constant. In the future studies, more levels of the effect size and the degree of unidimensionality could be included to provide a more complete picture about when the consequence of fitting unidimensional models to bifactor data on the general factor mean difference estimation becomes unacceptable.

Third, the study of Rhemtulla et al. (2012) indicated that the continuous approach (e.g., the MLR estimator) might not be appropriate for the ordinal data with asymmetrical threshold parameters. In this study, when exploring the impact of treating ordinal data as continuous on the general factor mean difference, all the threshold parameters were symmetrically distributed, and no obvious impact was found. In the future, to further explore this topic, the degree of symmetry for the threshold parameters might be manipulated.

Fourth, although item parameters in categorical CFAs and those in 2PL Models or GRMs can be converted to each other according to Equation 16, the shift in a given type of item parameter within one framework (e.g., CFA framework, as in this study) might lead to changes in different types of item parameters within the other framework (e.g., IRT framework). For example, for a given item, the DIF in the general factor loading for a bifactor CFA model corresponds to the main DIF in its general factor discrimination

parameter and also some changes in its specific factor discrimination parameter and item intercept for the corresponding bifactor GRM. Conversely, the DIF in the general factor discrimination parameter in a bifactor GRM corresponds to the main DIF in its general factor loading and also some changes in its specific factor loading and threshold parameter in the corresponding bifactor CFA. In the future study, the DIF can be manipulated in item parameters within the IRT framework to explore whether consistent conclusions regarding the general factor mean difference estimation could be made.

Fifth, the goodness of fit indices might provide supplemental information to significance tests in the process of evaluating differences in the general factor mean. In the current study, the means of several goodness of fit indices (i.e., CFI, WRMR/SRMR, RMSEA) were reported to provide some general sense of the sensitivity of these indices to the general factor mean difference under varied conditions. Incremental changes in these indices were not computed for each replicate dataset, which would be required to determine whether particular cutoffs are useful aids for decisions regarding the tenability of constraints.

Finally, to estimate the general factor mean difference in a bifactor model, at least one of the specific factor mean differences must be constrained to be zero, and others can be either freely estimated or constrained to be zero. In the current study, only the general factor mean difference was of interest, so all the specific factor means were constrained to be zero in both groups. However, in realistic situations, although researchers are usually most interested in the latent mean difference in the general factor when applying bifactor models, they might also want to estimate the specific factor mean differences at the same time. In these situations, they may freely estimate one or more specific factor mean

differences of interest. The choice of the specific factor(s) with the mean(s) constrained to be zero in the focal group might influence the latent mean difference estimation for both the general factor and other specific factors. Thus, the robustness of the general factor mean difference estimation to the choice of referent specific factors could be explored in the future.

Significance and Conclusions

Despite the prevalence and popularity of bifactor models, methodological issues in the estimation of multiple-group bifactor models have not been well studied. Different from the very few simulation studies regarding multiple-group bifactor models that focused on the DIF detection and item parameter recovery (e.g., Cai et al., 2011; Fukuhara & Kamata, 2011; Jeon et al., 2011), this study systematically explored factors that might influence the estimation and testing of the general factor mean difference for ordinal bifactor data.

In practice, ordinal bifactor data are often fitted with unidimensional models because consistent and absolute criteria to determine the degree of unidimensionality are lacking. Also, unidimensional models are preferred sometimes due to their theoretical simplicity, so researchers might assume a unidimensional data structure unless strong multidimensional evidence is found. In this study, I found that increase in the estimation bias of the general factor mean difference resulting from fitting unidimensional models to bifactor data was substantial (i.e., the absolute values of relative estimation bias were around 10-15%) when 66% of the explained common variance among items contributed to the general factor. However, the largest increase in squared estimation bias was still less than the decrease in estimated variance when fitting unidimensional models to bifactor data in the research conditions of the current study. Given the findings that the increase in

estimation bias grew by the same factor as the increase in effect size and that the estimation bias increased substantially as the degree of unidimensionality decreased, it would be expected that the unidimensional model might not be favored in terms of mean square error (i.e., MSE) as the absolute effect size of the general factor mean difference increases beyond 0.2 and the explained common variance among items contributed to the general factor decreases (i.e., less than 0.66).

According to previous simulation studies (e.g., Rhemtulla et al., 2012; Stark et al., 2006), the number of categories per item and sample size might influence whether it is appropriate to treat ordinal data as continuous. However, results of the current study revealed that the choice between the continuous approach (i.e., the MLR estimator) and the categorical approach (i.e., the WLSMV estimator) had a small influence on power and no obvious impact on estimation accuracy as long as there were no severe model misspecifications in the general factor mean difference estimation for the 3-point and 5-point scale data for total sample sizes of 600 and 1200.

Given that the DIF with relatively small magnitudes such as those in the current study may not be perfectly recovered in applied data analysis, the impact of constraining the noninvariant item parameters to be equal on the general factor mean difference estimation was also explored. It could be concluded from the current study that (1) when the effect size of the general factor mean difference was zero, the DIF in factor loadings had no impact on estimation for the general factor mean difference no matter whether the noninvariant loadings were freely estimated or constrained to be equal; (2) for the conditions with nonzero effect size of the general factor mean difference, ignoring the fact that some items did not discriminate examinees' performance in the general factor in the

focal group as well as what they did in the reference group would result in underestimation of the absolute difference in the general factor mean; (3) the estimation bias for the general factor mean difference resulting from setting equality constraints on the noninvariant general factor loadings increases by the same factor as that multiplied by the effect size or the magnitude of DIF; (4) the estimation of the general factor mean difference was somewhat robust to ignoring the DIF in general factor loadings in terms of estimation bias, given that the relative estimation bias was around 5% when the DIF= -0.15; (5) ignoring the differences in discrimination ability in the specific factors between groups would not bring in any bias in the estimation for the general factor mean difference, which means that the general factor mean difference estimation was completely robust to ignoring the DIF in specific factor loadings; (6) ignoring the fact that some items favored the reference group over the focal group would substantially overestimate the absolute value of the general factor mean difference, which means that the general factor mean difference estimation was not robust to the DIF in the threshold parameters; (7) for a given magnitude of DIF, the impact of constraining noninvariant threshold parameters on the estimation bias was similar across all effect sizes, including zero; and (8) for items with more categories (i.e., 5), freeing as many noninvariant threshold parameters as possible would somewhat reduce the estimation bias of the general factor mean difference.

In the current study, the dominant factors influencing the Type I error rate and power to detect the general factor mean difference were total sample size and effect size of the general factor mean difference. The Type I error rates fell in the designated limit of 0.025 to 0.075 in all conditions with no DIF or DIF in factor loadings, and they fell beyond this limit for many conditions with DIF in threshold parameters.

Finally, this study informs recommendations for applied researchers seeking to examine the general factor mean difference in ordinal, bifactor data. First, when multidimensionality is suspected but a unidimensional solution is preferred, it is important to examine the degree of unidimensionality and the effect size of the general factor mean difference. If the ECV is not very small (i.e., larger than 0.66) and the effect size is not very large (i.e., smaller than 0.2), it might be acceptable to fit the potential bifactor model with a unidimensional model. Second, if we are interested in the general factor mean difference for bifactor polytomous data, either the MLR estimator or the WLSMV estimator may be used, although it should be noted that there are not consistent opinions in the literature regarding whether these two estimators could detect DIF similarly (e.g., Desa, 2014; Meade & Lautenschlager, 2004; Stark et al., 2006). Third, DIF in threshold parameters cannot be completely detected using multiple-group categorical CFA models because one or two threshold parameters per item must be constrained to be equal to identify the model. Given that ignoring DIF in threshold parameters would yield substantial bias in the general factor mean difference estimate, I recommend examining the DIF in item intercept parameters within IRT framework before conducting the latent mean comparison of the general factor. If DIF is found in threshold parameters, it might not be appropriate to examine the general factor mean difference using multiple-group categorical CFA models because noninvariant threshold parameters may need to be constrained to be equal for identification purposes. Also, in this situation, the significance test of the general factor mean difference might not be reliable due to potentially large estimation bias. Fourth, given that the general factor difference estimation was somewhat robust to ignoring DIF in general factor loadings and completely robust to ignoring DIF in specific factor loadings,

constraining the general factor loadings with smaller DIF (i.e., less than 0.15) or constraining the specific factor loadings with any size of DIF might be acceptable if researchers want to estimate the general factor mean difference based on information from more observed variables.

Reference

Ackerman, T. A. (2005). Multidimensional item response theory models. *Wiley StatsRef: Statistics Reference Online*.

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1-23.

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society. Series B (Methodological)*, 283-301.

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness‐of‐fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, *36*(4), 277-300.

Asparouhov, T., Muthén, B., & Muthén, B. O. (2006). Robust chi square difference testing with mean and variance adjusted test statistics. *matrix*, *1*(5).

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508.

Babcock, B. (2011). Estimating a noncompensatory IRT model using Metropolis within Gibbs sampling. *Applied Psychological Measurement*, *35*(4), 317-329.

Berkeljon, A. (2012). Multidimensional item response theory in clinical measurement: A bifactor graded-response model analysis of the outcome-questionnaire-45.2 (doctoral dissertation). Brigham Young University, Provo, United States.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education*, *15*(2), 113-141.

Bolt, D. M., Hare, R. D., Vitale, J. E., & Newman, J. P. (2004). A Multigroup Item Response Theory Analysis of the Psychopathy Checklist-Revised. *Psychological Assessment*, *16*(2), 155-168.

Borsboom, D. (2006). When does measurement invariance matter? *Medical care*, *44*(11), S176-S181.

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

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581-612.

Cai, L., Du Toit, S. H. C., & Thissen, D. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. *Chicago, IL: Scientific Software International*.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221-248.

Chang, H. H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure. *ETS Research Report Series, 1995*(1), i-30.

Chang, Y. F. (2015). A Restricted Bifactor Model of Subdomain Relative Strengths and Weaknesses (Doctoral dissertation). University of Minnesota, Urban, United States.

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.

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

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

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*(1), 219-251.

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

Cohen, A. S., Kim, S. H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20*(1), 15-26.

De Beuckelaer, A., & Swinnen, G. (2018). Biased latent variable mean comparisons due to measurement noninvariance: A simulation study. In *Cross-Cultural Analysis* (pp. 163-192). Routledge.

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.

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354-378.

Desa, M., & Deana, Z. N. (2012). Bifactor multidimensional item response theory modeling for subscores estimation, reliability, and classification (Doctoral dissertation). University of Kansas, Lawrence, United States.

Desa, D. (2014). Evaluating measurement invariance of TALIS 2013 complex scales.

De Ayala, R. J. (2009). The Theory and Practice of Item Response Theory/RJ de Ayala. *New York: Guilford Publications Incorporated*.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of educational measurement*, *23*(4), 355-368.

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*(4), 278-295.

Fleer, P. F., Raju, N. S., & van der Linden, W. J. (1995). A Monte Carlo assessment of DFIT with dichotomously scored unidimensional tests. In *Annual Meeting of the American Educational Research Association, San Francisco*.

Flowers, C. P., Raju, N. S. & Oshima, T. C. (2002). A comparison of measurement equivalence methods based on confirmatory factor analysis and item response Theory. In *National Council on Measurement in Education (NCME) Annual Meeting, New Orleans, Los Angeles*.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622.

Gelman, A. (1993). Iterative and non-iterative simulation algorithms. *Computing science and statistics*, 433-433.

Geman, S., & Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics*, *20*(5-6), 25-62.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika*, *57*(3), 423-436.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4-19.

Gignac, G. E., & Watkins, M. W. (2013). Bifactor modeling and the estimation of model-based reliability in the WAIS-IV. *Multivariate Behavioral Research*, *48*(5), 639-662.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407-434.

Haley, S. M., Ni, P., Dumas, H. M., Fragala-Pinkham, M. A., Hambleton, R. K., Montpetit, K. & Tucker, C. A. (2009). Measuring global physical health in children with cerebral palsy: illustration of a multidimensional bifactor model and computerized adaptive testing. *Quality of Life Research*, *18*(3), 359-370.

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*, *7*(4), 534-556.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97-109.

Holzinger, K. J., & Swineford, F. (1937). The bifactor method. *Psychometrika*, *2*(1), 41-54.

Hong, S., Malik, M. L., & Lee, M. K. (2003). Testing configural, metric, scalar, and latent mean invariance across genders in sociotropy and autonomy using a non-Western sample. *Educational and psychological measurement*, *63*(4), 636-654.

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.

Jones, R. N., & Gallo, J. J. (2002). Education and sex differences in the mini-mental state examination: effects of differential item functioning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*(6), 548-558.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*(1), 136-153.

Kim, S. H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of educational measurement*, *29*(1), 51-66.

Kim, S. H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement*, *22*(4), 345-355.

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

Kline, R. B. (2011). Principles and practice of structural equation modeling. 3rd edition. New York: The Guilford Press.

Kogar, H. (2018). An Examination of Parametric and Nonparametric Dimensionality Assessment Methods with Exploratory and Confirmatory Mode. *Journal of Education and Learning*, *7*(3), 148.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality assessment using the full-information item bifactor analysis for graded response data: An illustration with the State Metacognitive Inventory. *Educational and Psychological Measurement*, *68*(4), 695-709.

Li, Y. (2011). Exploring the full-information bifactor model in vertical scaling with construct shift (Unpublished doctoral dissertation). University of Maryland, College Park, United States.

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

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: a unified framework. *Journal of the American Statistical Association*, *100*(471), 1009-1020.

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.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International journal of educational research*, *13*(2), 127-143.

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*(2), 289-311.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical care*, *44*(11), S69-S77.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087-1092.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515.

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461-473.

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

Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*(4), 453-477.

Morales, L. S., Flowers, C., Gutierrez, P., Kleinman, M., & Teresi, J. A. (2006). Item and scale differential functioning of the Mini-Mental State Exam assessed using the differential item and test functioning (DFIT) framework. *Medical care*, *44*(11 Suppl 3), S143.

Muthén, L. K., & Muthén, B. O. (1998). Statistical analysis with latent variables. *Mplus User's guide*, *2012*.

Muthén, B. O. (1998–2004). Mplus technical appendices. Los Angeles, CA: Muthén & Muthén.

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257-274.

Oishi, S. (2006). The concept of life satisfaction across cultures: An IRT analysis. *Journal of Research in Personality*, *40*(4), 411-423.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and psychological measurement*, *53*(2), 301-314.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*(4), 401-412.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*(2), 133-144.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological bulletin, 114*(3), 552.

Reise, S. P., & Henson, J. M. (2003). A discussion of modern versus traditional psychometrics as applied to personality assessment scales. *Journal of personality assessment, 81*(2), 93-103.

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., 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. (2012). The rediscovery of bifactor measurement models. *Multivariate behavioral research, 47*(5), 667-696.

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.

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi‐factor, the testlet, and a second‐order multidimensional IRT model. *Journal of Educational Measurement, 47*(3), 361-372.

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

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222.

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.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the Multidimensional Personality Questionnaire Stress Reaction scale. *Journal of Personality and Social Psychology, 75*(5), 1350-1362.

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

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-90.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*(4), 589-617.

Stout, W., Douglas, B., Junker, B., & Roussos, L. (1999). Dimtest. *Computer software, The William Stout Institute for Measurement, Champaigne, IL*.

Sweeney, K. P. (1996). A Monte Carlo investigation of the likelihood-ratio procedure in the detection of differential item functioning (Doctoral dissertation). Fordham University, Bronx, United States.

Sympson, J. B. (1978). A model for testing with multidimensional items. In *Proceedings of the 1977 computerized adaptive testing conference* (No. 00014). Minneapolis, MN: University of Minneapolis, Department of Psychology, Psychometric Methods Program.

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

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*(1), 3-46.

Teresi, J. A. (2006). Overview of quantitative measurement methods: Equivalence, invariance, and differential item functioning in health applications. *Medical Care*, *44*(11), S39-S49.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple - categorical - response models. *Journal of Educational Measurement*, *26*(3), 247-260.

Wainer, H. (2000). CATs: Whither and whence. *ETS Research Report Series*, *2000*(2).

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer, Dordrecht.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement, 27*(6), 479-498.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58.

Woods, C. M., Cai, L., & Wang, M. (2012). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.

Yang, Y. (2008). *Partial invariance in loadings and intercepts—Their interplays and implications for latent mean comparisons.* (Unpublished doctoral dissertation). The University of Nebraska, Lincoln.

Yang, F. M., Tommet, D., & Jones, R. N. (2009). Disparities in self-reported geriatric depressive symptoms due to sociodemographic differences: An extension of the bifactor item response theory model for use in differential item functioning. *Journal of Psychiatric Research*, *43*(12), 1025-1035.

Yu, C. Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes (Unpublished doctoral dissertation). University of California Los Angeles, Los Angeles, CA.

Yu, C.-Y., & Muthen, B. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. In annual meeting of the American Educational Research Association, New Orleans, LA.

APPENDIX A

PREVIOUS SIMULATION STUDYS REGARDING DIF DETECTION

Table 1A

Simulation Studies on Factors Influencing Type I Error Rates in Detecting DIF

Reference	Data Generation Factors	Data Analysis	Main Results
Cohen, Kim & Wollack (1996)	a) model type (2PL or 3PL IRT model); b) sample size	a) correctly specifying the models; b) misspecifying the 3PL model by fixing the pseudo-guessing parameter to the average value of the pseudo-guessing parameters of all the items	a) the type I error rates were close to the nominal alpha level for the 2PL model conditions; b) the type I error rates were inflated for both correctly specified 3PL models and misspecified 3PL models, especially when the nominal alpha level was at .0005 to .005; c) sample size does not influence type I error rates
Bolt (2002)	a) model type (GRM or alternative models to GRM); b) sample size; c) the latent mean difference in ability	applying different DIF detection methods (LR-GRM, DFIT-GRM or Poly-SIBTEST)	a) slight misspecification of the model would lead to large inflation of Type I error rates when applying the LR-GRM, and such inflation was especially severe when the sample size was large; b) there were less Type I error rates inflation due to model misspecification if using DFIT-GRM; c) Type I error rates were unaffected by the generating models using Poly-SIBTEST
Ankenmann, Witt & Dunbar (1999)	a) sample size; b) the latent mean difference in ability; c) parameter values for the studied item	applying different DIF detection methods (LR tests or Mantel procedure)	a) both LR tests and Mantel procedure showed good control over Type I error rates when the distributions of ability parameters were identical across groups ; b) when the latent mean difference in ability was nonzero, LR tests still maintained acceptable control over Type I error rates whereas the Mantel procedure lacked control over Type I error rates, and the inflation got worse for larger sample size and higher discrimination parameter values

Table 1A Continued

Reference	Data Generation Factors	Data Analysis	Main Results
Stark, Chernyshenko & Drasgow (2006)	a) amount of DIF; b) Type of DIF when present; c) the latent mean difference in ability; d) number of response categories; e) sample size	a) applying different DIF detection methods (IRT-LR or chi-square difference test under the traditional CFA); b) applying different baseline model (free baseline model or constrained baseline model); c) applying different criterion of the *p* value (.05 or Bonferroni corrected)	a) when using the constrained-baseline model, both IRT-LR and chi-square difference tests under CFA showed substantial Type I error inflation unless no DIF existed in the fully constrained model; b) the Type I error inflation could be reduced by applying Bonferroni corrected critical *p* value; c) larger sample size was slightly related to larger Type I error rates; d) the latent mean difference in ability did not substantially influence the Type I error rate
Wang & Yeh (2003)	a) model type (2PL, 3PL or GRM); b) percentage of DIF; c) DIF direction (one sided or both sided)	applying LR tests with different anchor item methods (all-other, 1-item constant, 4-item constant or 10-item constant)	a) when conducting LR tests using all other items as anchor, Type I error inflation occurred when the percentage of items with DIF reached 12% under the 3PL model and 20% under the 2PL model and the GRM for the conditions in one-side conditions; b) the performance of the constrained baseline model in controlling over Type I error rates was determined by average signed area; c) all the three constant methods had good control over Type I error rates

Table 2A

Simulation Studies on the Factors Influencing Power in Detecting DIF

Reference	Data Generation Factors	Data Analysis	Main Results
Sweeney (1996)	a) item parameter values; b) amount of DIF; c) latent mean difference in ability; d) ratio of sample size between groups	applying LR tests for all the conditions	a) the item with larger effect size of DIF was more easily detected as showing DIF; b) the magnitude of item parameters influenced the power to detect DIF for them; c) for a given total sample size, the power to detect DIF was higher for equal sample size conditions than the conditions with much fewer examinees in the focal group ; d) the power to detect DIF depended on the differences between the IRFs for the reference group and the IRFs for the focal group and the number of focal group examinees located on the latent ability continuum where the IRFs differ across groups
Bolt (2002)	a) model type (GRM or alternative models to GRM); b) sample size; c) the latent mean difference in ability	applying different DIF detection methods (LR-GRM, DFIT-GRM or Poly-SIBTEST)	a) sample size was the main factor influencing power; b) power was not influenced by the generating models obviously; c) there was a slight reduction in power when the latent mean difference was nonzero if using DFIT-GRM; d) in comparison with Poly-SIBTEST, LR-GRM and DFIT-GRM showed greater power in detecting DIF
Ankenmann, Witt & Dunbar (1999)	a) sample size; b) the latent mean difference in ability; c) parameter values for the studied item; d) the pattern of DIF in threshold parameters	a) applying different DIF detection methods (LR tests or Mantel procedure)	a) the power was influenced by the sample size for both of the methods; b) the power was higher for larger discrimination parameter conditions; c) Mantel procedure showed greater power than LR tests for the constant DIF pattern conditions when the person ability distributions were identical across groups; d) for the balanced DIF pattern conditions, LR tests showed much higher power than Mantel procedure

Table 2A Continued

Reference	Data Generation Factors	Data Analysis	Main Results
Stark, Chernyshenko & Drasgow (2006)	a) amount of DIF; b) Type of DIF when present; c) the latent mean difference in ability; d) number of response categories; e) sample size	a) applying different DIF detection methods (IRT-LR or chi-square difference test under the traditional CFA); b) applying different baseline model (free baseline model or constrained baseline model); c) applying different criterion of the *p* value (.05 or Bonferroni corrected)	a) perfect detection was achieved for all large DIF conditions; b) sample sizes, analysis methods and baseline models influenced power for small DIF conditions; c) Bonferroni corrected critical *p* value reduced power as well; d) the increase in the number of categories improved accuracy of DIF detection using traditional CFA models; e) for small sample sizes, traditional CFA models performed better than IRT in power to detect DIF; f) the free-baseline models performed better than the constrained-baseline models for both LR tests under IRT models and chi-squire difference tests under CFA models
Wang & Yeh (2003)	a) model type (2PL, 3PL or GRM); b) percentage of DIF; c) DIF direction (one sided or both sided)	applying LR tests with different anchor item methods (all-other, 1-item constant, 4-item constant or 10-item constant)	a) using 1 anchor items could show acceptable power ; b) using 4 or 10 anchor items led to higher power

APPENDIX B

DETAILED RESULTS OF THE SIMULATION STUDY

Table B1

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.040	.0010	.007	.961	.032	-.0021	.006	.969
	1200	---	---	---	---	.041	.0009	.003	.960	.035	.0031	.003	.965
Unidimensional model with WLSMV estimator	600	.061	-.0003	.007	.963	.040	.0012	.007	.960	.051	-.0020	.006	.968
	1200	.063	.0044	.004	.963	.040	.0008	.003	.959	.051	.0032	.003	.965
Bifactor model with MLR estimator	600	---	---	---	---	.039	.0008	.007	.960	.033	-.0031	.006	.969
	1200	---	---	---	---	.039	.0009	.004	.961	.036	.0034	.003	.965
Bifactor model with WLSMV estimator	600	.046	-.0012	.008	.968	.042	.0013	.007	.959	.046	-.0022	.006	.971
	1200	.068	.0043	.004	.962	.040	.0009	.004	.961	.045	.0032	.003	.965

Note: N denotes total sample size. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B2

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.041	-.0023	.007	.960	.043	-.0027	.006	.961
	1200	---	---	---	---	.042	-.0006	.003	.957	.039	.0019	.003	.961
Unidimensional model with WLSMV estimator	600	.066	-.0035	.007	.960	.051	-.0023	.006	.964	.054	-.0024	.006	.959
	1200	.059	.0024	.004	.959	.069	-.0006	.003	.957	.057	.0018	.003	.961
Bifactor model with MLR estimator	600	---	---	---	---	.038	-.0024	.008	.962	.041	-.0029	.007	.960
	1200	---	---	---	---	.041	-.0005	.004	.961	.040	.0022	.004	.962
Bifactor model with WLSMV estimator	600	.060	-.0041	.009	.961	.057	-.0026	.008	.964	.057	-.0027	.007	.960
	1200	.056	.0027	.004	.958	.067	-.0006	.004	.956	.054	.0020	.004	.961

Note: N denotes total sample size. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B3

Power (Δκ = -0.10), Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.188	.0048 (-4.777%)	.007	.961	.185	.0022 (-2.177%)	.006	.967
	1200	---	---	---	---	.339	.0047 (-4.714%)	.003	.970	.355	.0036 (-3.609%)	.003	.954
Unidimensional model with WLSMV estimator	600	.232	.0007 (-0.651%)	.008	.956	.191	.0044 (-4.398%)	.007	.960	.217	.0019 (-1.862%)	.006	.967
	1200	.369	.0048 (-4.839%)	.004	.965	.336	.0042 (-4.164%)	.003	.970	.405	.0031 (-3.137%)	.003	.954
Bifactor model with MLR estimator	600	---	---	---	---	.183	.0035 (-3.542%)	.007	.961	.187	-.0002 (0.186%)	.007	.967
	1200	---	---	---	---	.347	.0020 (-2.040%)	.003	.971	.355	.0009 (-0.905%)	.004	.956
Bifactor model with WLSMV estimator	600	.220	-.0015 (1.492%)	.008	.958	.187	.0045 (-4.536%)	.007	.956	.229	-.0008 (0.845%)	.007	.969
	1200	.363	.0015 (-1.506%)	.004	.964	.337	.0013 (-1.325%)	.003	.971	.403	.0002 (-0.180%)	.004	.957

Note: $\Delta\kappa$ = *-0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. *N* denotes total sample size. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B4

Power (Δκ = -0.10), Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.172	.0094 (-9.354%)	.006	.968	.169	.0122 (-12.169%)	.007	.956
	1200	---	---	---	---	.335	.0069 (-6.914%)	.004	.953	.321	.0086 (-8.610%)	.003	.975
Unidimensional model with WLSMV estimator	600	.219	.0074 (-7.357%)	.007	.951	.206	.0117 (-11.686%)	.006	.969	.166	.0156 (-15.606%)	.006	.955
	1200	.378	.0101 (-10.123%)	.004	.957	.385	.0094 (-9.372%)	.004	.952	.327	.0118 (-11.789%)	.003	.973
Bifactor model with MLR estimator	600	---	---	---	---	.170	.0025 (-2.507)	.008	.973	.165	.0054 (-5.413%)	.008	.958
	1200	---	---	---	---	.339	-.0002 (0.196%)	.004	.957	.320	.0018 (-1.810%)	.004	.977
Bifactor model with WLSMV estimator	600	.210	-.0023 (2.324%)	.009	.951	.205	.0019 (-1.875%)	.008	.971	.166	.0055 (-5.526%)	.008	.959
	1200	.358	.0005 (-0.506%)	.004	.971	.383	-.0008 (0.845%)	.004	.955	.325	.0013 (-1.341%)	.004	.977

Note: $\Delta\kappa$ = *-0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. *N* denotes total sample size. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B5

Power (Δκ = -0.20), Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.607	.0096 (-4.819%)	.007	.953	.612	.0058 (-2.885%)	.007	.961
	1200	---	---	---	---	.892	.0063 (-3.131%)	.003	.955	.908	.0067 (-3.374%)	.003	.961
Unidimensional model with WLSMV estimator	600	.646	.0052 (-2.585%)	.008	.962	.606	.0081 (-4.037%)	.007	.950	.668	.0053 (-2.629%)	.007	.962
	1200	.888	.0038 (-1.917%)	.004	.962	.884	.0046 (-2.305%)	.004	.959	.933	.0057 (-2.868%)	.003	.963
Bifactor model with MLR estimator	600	---	---	---	---	.592	.0041 (-2.055%)	.007	.955	.594	.0012 (-0.603%)	.007	.965
	1200	---	---	---	---	.886	.0012 (-0.575%)	.004	.957	.908	.0016 (-0.777%)	.003	.964
Bifactor model with WLSMV estimator	600	.646	.0011 (-0.570%)	.007	.964	.605	.0025 (-1.250%)	.007	.950	.672	-.0004 (0.194%)	.007	.963
	1200	.895	-.0019 (0.962%)	.004	.958	.884	-.0011 (0.559%)	.004	.955	.929	-.0001 (0.032%)	.004	.967

Note: $\Delta\kappa$ = *-0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. *N* denotes total sample size. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B6

Power (Δκ = -0.20), Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF

Data analysis models	*N*	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
Unidimensional model with MLR estimator	600	---	---	---	---	.560	.0229 (-11.436%)	.007	.951	.568	.0185 (-9.271%)	.007	.954
	1200	---	---	---	---	.868	.0152 (-7.609%)	.003	.953	.897	.0125 (-6.238%)	.003	.962
Unidimensional model with WLSMV estimator	600	.604	.0176 (-8.784%)	.007	.957	.547	.0275 (-13.745%)	.006	.945	.620	.0251 (-12.526%)	.006	.948
	1200	.888	.0175 (-8.755%)	.003	.963	.864	.0198 (-9.895%)	.003	.947	.913	.0190 (-9.491%)	.003	.952
Bifactor model with MLR estimator	600	---	---	---	---	.549	.0097 (-4.826%)	.008	.958	.564	.0052 (-2.617%)	.008	.964
	1200	---	---	---	---	.863	.0012 (-0.594%)	.004	.962	.897	-.0013 (0.655%)	.003	.974
Bifactor model with WLSMV estimator	600	.597	-.0010 (0.496%)	.008	.964	.548	.0081 (-4.055%)	.008	.964	.627	.0043 (-2.173%)	.008	.962
	1200	.886	-.0017 (0.850%)	.004	.964	.865	-.0003 (0.133%)	.004	.960	.916	-.0024 (1.197%)	.003	.975

Note: $\Delta\kappa$ = *-0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. *N* denotes total sample size. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B7

Goodness of Fit Indices for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = 0)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.952/.952	.046/.047	.054/.053	.946/.946	.046/.047	.062/.062
	1200	---	---	---	.953/.953	.039/.039	.053/.053	.947/.947	.040/.040	.061/.061
Unidimensional model with WLSMV estimator	600	.984/.986	1.229/1.250	.042/.039	.980/.986	1.291/1.322	.053/.044	.978/.988	1.303/1.343	.056/.040
	1200	.983/.986	1.440/1.459	.044/.040	.979/.986	1.572/1.597	.055/.044	.977/.988	1.599/1.632	.056/.040
Bifactor model with MLR estimator	600	---	---	---	.996/.996	.033/.034	.012/.012	.997/.997	.032/.033	.012/.012
	1200	---	---	---	.998/.998	.023/.024	.008/.008	.999/.999	.022/.023	.008/.008
Bifactor model with WLSMV estimator	600	.999/.999	1.013/1.039	.007/.007	.999/.999	.845/.890	.010/.008	.998/.999	.844/.902	.012/.007
	1200	.999/.999	1.023/1.050	.006/.006	.999/.999	.852/.896	.008/.006	.999/.999	.852/.911	.010/.006

Note: *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B8

Goodness of Fit Indices for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = 0)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.750/.749	.091/.091	.146/.145	.728/.728	.096/.096	.169/.168
	1200	---	---	---	.750/.750	.088/.088	.145/.145	.730/.730	.093/.093	.168/.167
Unidimensional model with WLSMV estimator	600	.912/.922	2.639/2.651	.122/.114	.898/.922	3.122/3.136	.151/.132	.892/.935	3.262/3.280	.159/.123
	1200	.907/.918	3.599/3.608	.125/.116	.893/.919	4.314/4.325	.154/.133	.887/.933	4.501/4.515	.161/.123
Bifactor model with MLR estimator	600	---	---	---	.997/.997	.031/.033	.012/.012	.998/.998	.030/.031	.012/.012
	1200	---	---	---	.999/.999	.022/.024	.008/.008	.999/.999	.021/.023	.009/.009
Bifactor model with WLSMV estimator	600	.999/.999	.972/1.003	.009/.009	.999/.999	.805/.854	.012/.009	.999/.999	.807/.874	.013/.009
	1200	.999/.999	.987/1.017	.008/.007	.999/.999	.826/.877	.011/.008	.999/1.000	.841/.908	.013/.008

Note: *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B9

Goodness of Fit Indices for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = -0.10)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.951/.950	.046/.048	.054/.054	.946/.946	.046/.049	.062/.062
	1200	---	---	---	.952/.952	.039/.041	.053/.053	.948/.947	.039/.042	.061/.061
Unidimensional model with WLSMV estimator	600	.984/.984	1.225/1.277	.042/.041	.979/.984	1.297/1.365	.054/.047	.978/.986	1.301/1.399	.055/.043
	1200	.983/.984	1.442/1.507	.044/.042	.979/.984	1.578/1.666	.055/.047	.977/.986	1.593/1.723	.056/.043
Bifactor model with MLR estimator	600	---	---	---	.996/.996	.033/.036	.012/.013	.997/.997	.032/.035	.011/.012
	1200	---	---	---	.998/.998	.024/.027	.008/.009	.998/.998	.022/.026	.008/.009
Bifactor model with WLSMV estimator	600	.999/.998	1.012/1.074	.008/.011	.999/.998	.845/.942	.010/.014	.998/.997	.845/.983	.012/.015
	1200	.999/.998	1.023/1.113	.007/.011	.999/.998	.857/1.005	.009/.015	.999/.998	.854/1.070	.010/.016

Note: *Δκ =- 0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B10

Goodness of Fit Indices for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = -0.10)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.751/.750	.091/.092	.145/.145	.728/.728	.096/.097	.169/.168
	1200	---	---	---	.750/.750	.088/.089	.145/.145	.729/.728	.093/.094	.168/.168
Unidimensional model with WLSMV estimator	600	.912/.921	2.647/2.673	.122/.115	.898/.921	3.109/3.140	.151/.132	.891/.933	3.257/3.298	.159/.124
	1200	.907/.917	3.589/3.617	.124/.117	.893/.918	4.312/4.350	.154/.134	.887/.932	4.517/4.565	.161/.125
Bifactor model with MLR estimator	600	---	---	---	.997/.997	.031/.034	.012/.013	.998/.998	.029/.033	.012/.012
	1200	---	---	---	.999/.998	.022/.026	.008/.009	.999/.999	.021/.025	.009/.010
Bifactor model with WLSMV estimator	600	.999/.998	.978/1.044	.010/.013	.999/.998	.806/.910	.012/.015	.999/.998	.802/.944	.012/.014
	1200	.999/.999	.981/1.076	.007/.012	.999/.998	.824/.990	.011/.017	.999/.998	.836/1.053	.013/.017

Note: *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B11

Goodness of Fit Indices for the Weak Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = -0.20)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.952/.950	.046/.052	.053/.054	.947/.945	.046/.053	.062/.062
	1200	---	---	---	.953/.951	.039/.045	.053/.054	.948/.946	.040/.047	.061/.062
Unidimensional model with WLSMV estimator	600	.984/.980	1.227/1.353	.042/.047	.980/.979	1.290/1.468	.053/.054	.978/.980	1.299/1.558	.055/.052
	1200	.983/.979	1.437/1.638	.043/.048	.979/.978	1.571/1.849	.054/.055	.977/.980	1.596/1.981	.056/.052
Bifactor model with MLR estimator	600	---	---	---	.996/.995	.033/.041	.012/.015	.997/.996	.032/.041	.011/.015
	1200	---	---	---	.998/.997	.023/.033	.008/.012	.998/.997	.022/.033	.008/.012
Bifactor model with WLSMV estimator	600	.999/.994	1.019/1.169	.008/.022	.999/.992	.847/1.093	.010/.030	.998/.991	.843/1.196	.011/.032
	1200	.999/.994	1.020/1.292	.006/.024	.999/.992	.852/1.286	.008/.033	.999/.991	.853/1.439	.010/.034

Note: *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B12

Goodness of Fit Indices for the Strong Specific Factor Conditions for the Generated Multiple-group Bifactor IRT Models without DIF (Δκ = -0.20)

Data analysis models	*N*	Binary data			3-point scale data			5-point scale data		
		CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA	CFI	SRMR (WRMR)	RMSEA
Unidimensional model with MLR estimator	600	---	---	---	.750/.748	.091/.094	.145/.145	.727/.726	.096/.099	.169/.168
	1200	---	---	---	.751/.750	.088/.091	.145/.145	.729/.728	.093/.096	.168/.168
Unidimensional model with WLSMV estimator	600	.912/.918	2.641/2.705	.122/.117	.898/.918	3.119/3.196	.151/.135	.891/.930	3.260/3.371	.159/.127
	1200	.907/.914	3.588/3.673	.124/.119	.894/.915	4.298/4.410	.153/.136	.887/.928	4.508/4.665	.161/.128
Bifactor model with MLR estimator	600	---	---	---	.997/.996	.031/.039	.012/.015	.998/.997	.029/.039	.012/.014
	1200	---	---	---	.999/.998	.023/.033	.009/.013	.999/.998	.021/.033	.008/.013
Bifactor model with WLSMV estimator	600	.999/.996	.969/1.129	.009/.022	.999/.995	.808/1.055	.012/.030	.999/.995	.806/1.159	.013/.031
	1200	.999/.996	.981/1.256	.008/.025	.999/.995	.828/1.270	.012/.033	.999/.995	.834/1.442	.013/.034

Note: *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0. SRMRs were given for models estimated with the MLR estimator, and WRMRs were given for models estimated with the WLSMV estimator.

Table B13

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.056	-.0007	.008	.965	.075	-.0001	.009	.949	.033	.0027	.008	.967
	Yes	.051	-.0006	.008	.966	.067	-.0002	.008	.950	.033	.0026	.008	.966
-0.10	No	.051	-.0034	.009	.958	.060	-.0022	.009	.955	.030	-.0014	.008	.970
	Yes	.054	-.0032	.008	.958	.063	-.0022	.008	.956	.030	-.0014	.007	.970
-0.15	No	.057	-.0030	.009	.960	.072	-.0016	.009	.957	.043	.0016	.008	.959
	Yes	.056	-.0027	.008	.960	.067	-.0015	.008	.954	.041	.0016	.007	.959

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B14

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.066	.0048	.005	.951	.056	-.0036	.004	.967	.040	-.0018	.004	.961
	Yes	.063	.0047	.004	.952	.055	-.0034	.004	.967	.041	-.0018	.003	.960
-0.10	No	.048	.0010	.004	.966	.058	-.0037	.004	.960	.034	.0002	.004	.967
	Yes	.041	.0009	.004	.966	.067	-.0035	.004	.959	.034	.0001	.004	.967
-0.15	No	.056	.0005	.004	.964	.056	-.0026	.004	.963	.028	.0026	.004	.972
	Yes	.051	.0005	.004	.963	.056	-.0025	.004	.965	.029	.0026	.003	.971

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B15

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.204	-.0023 (2.264%)	.009	.958	.202	-.0011 (1.115%)	.007	.969	.178	-.0016 (1.595%)	.008	.963
	Yes	.211	-.0006 (0.586%)	.008	.960	.202	.0007 (-0.726%)	.007	.971	.180	.0002 (-0.226%)	.008	.963
-0.10	No	.193	-.0007 (0.686%)	.008	.966	.222	-.0024 (2.377%)	.008	.960	.190	-.0028 (2.803%)	.008	.971
	Yes	.199	.0026 (-2.551%)	.008	.971	.220	.0015 (-1.498%)	.008	.959	.188	.0009 (-0.919%)	.007	.972
-0.15	No	.181	-.0010 (0.981%)	.008	.962	.218	.0029 (-2.917%)	.009	.958	.168	.0025 (-2.489%)	.008	.972
	Yes	.184	.0038 (-3.805%)	.008	.963	.217	.0084 (-8.437%)	.008	.960	.168	.0080 (-7.970%)	.007	.971

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B16

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.347	-.0010 (1.040%)	.004	.961	.376	.0006 (-0.624%)	.004	.965	.294	.0030 (-3.028%)	.004	.971
	Yes	.343	.0006 (-0.585%)	.004	.960	.381	.0024 (-2.399%)	.004	.966	.297	.0048 (-4.844%)	.004	.970
-0.10	No	.348	-.0013 (1.338%)	.004	.958	.376	.0007 (-0.655%)	.004	.959	.336	-.0044 (4.363%)	.004	.960
	Yes	.342	.0021 (-2.064%)	.004	.958	.371	.0044 (-4.416%)	.004	.958	.335	-.0004 (0.429%)	.004	.962
-0.15	No	.346	-.0016 (1.582%)	.004	.964	.373	.0012 (-1.194%)	.004	.957	.318	.0008 (-0.825%)	.004	.963
	Yes	.335	.0032 (-3.180%)	.004	.963	.362	.0067 (-6.713%)	.004	.951	.314	.0063 (-6.281%)	.004	.959

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B17

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.569	.0049 (-2.461%)	.009	.963	.631	-.0002 (0.081%)	.008	.967	.612	-.0025 (1.247%)	.008	.961
	Yes	.558	.0083 (-4.129%)	.008	.961	.622	.0032 (-1.620%)	.008	.966	.608	.0011 (-0.562%)	.008	.963
-0.10	No	.599	-.0028 (1.396%)	.008	.967	.637	-.0011 (0.567%)	.008	.967	.571	.0029 (-1.433%)	.008	.955
	Yes	.593	.0042 (-2.091%)	.008	.968	.630	.0063 (-3.142%)	.007	.962	.570	.0099 (-4.970%)	.008	.957
-0.15	No	.580	.0012 (-0.608%)	.009	.973	.604	.0007 (-0.365%)	.009	.955	.558	-.0011 (0.536%)	.008	.962
	Yes	.561	.0103 (-5.165%)	.008	.971	.590	.0119 (-5.960%)	.008	.959	.556	.0100 (-4.997%)	.007	.962

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B18

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.856	.0028 (-1.410%)	.004	.960	.882	.0005 (-0.256%)	.004	.956	.860	.0018 (-0.903%)	.004	.960
	Yes	.854	.0062 (-3.124%)	.004	.960	.881	.0041 (-2.046%)	.004	.951	.862	.0055 (-2.731%)	.004	.957
-0.10	No	.858	-.0006 (0.298%)	.004	.968	.874	.0027 (-1.341%)	.004	.954	.882	-.0031 (1.530%)	.004	.953
	Yes	.865	.0060 (-2.981%)	.004	.966	.872	.0099 (-4.946%)	.004	.950	.877	.0044 (-2.195%)	.004	.955
-0.15	No	.829	.0043 (-2.160%)	.004	.965	.863	.0030 (-1.478%)	.005	.951	.882	-.0010 (0.484%)	.004	.961
	Yes	.822	.0136 (-6.801%)	.004	.966	.854	.0141 (-7.043%)	.004	.947	.882	.0102 (-5.122%)	.004	.957

Note: N denotes total sample size. $\Delta\kappa$ = *-0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B19

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.042	-.0044	.007	.971	.058	.0001	.008	.961	.050	.0001	.008	.952
	Yes	.040	-.0044	.007	.970	.053	.0001	.008	.960	.050	.0001	.008	.951
-0.10	No	.058	-.0012	.008	.963	.058	-.0058	.008	.957	.038	.0027	.008	.963
	Yes	.054	-.0012	.008	.963	.061	-.0058	.008	.957	.038	.0027	.008	.963
-0.15	No	.064	.0002	.009	.957	.050	.0068	.008	.967	.039	-.0038	.007	.962
	Yes	.068	.0002	.009	.957	.057	.0067	.008	.967	.039	-.0038	.007	.962

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B20

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.059	.0021	.004	.960	.064	.0025	.004	.956	.038	.0014	.004	.963
	Yes	.065	.0021	.004	.960	.061	.0025	.004	.956	.038	.0014	.004	.963
-0.10	No	.047	-.0001	.004	.967	.066	-.0003	.004	.957	.040	.0016	.004	.960
	Yes	.053	-.0001	.004	.967	.063	-.0003	.004	.957	.039	.0016	.004	.960
-0.15	No	.065	-.0010	.004	.951	.046	-.0005	.004	.968	.035	.0010	.004	.967
	Yes	.062	-.0010	.004	.950	.049	-.0005	.004	.968	.034	.0010	.004	.967

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B21

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.226	-.0048 (4.772%)	.009	.956	.217	.0020 (-1.977%)	.008	.960	.184	-.0033 (3.289%)	.007	.965
	Yes	.223	-.0048 (4.752%)	.009	.956	.214	.0020 (-2.015%)	.008	.960	.184	-.0033 (3.286%)	.007	.965
-0.10	No	.176	.0054 (-5.446%)	.008	.969	.215	.0010 (-0.957%)	.008	.964	.178	-.0029 (2.899%)	.007	.961
	Yes	.176	.0055 (-5.481%)	.008	.969	.215	.0010 (-1.002%)	.008	.964	.178	-.0028 (2.844%)	.007	.961
-0.15	No	.173	.0064 (-6.414%)	.008	.970	.227	-.0007 (0.653%)	.008	.964	.188	-.0025 (2.463%)	.007	.967
	Yes	.171	.0066 (-6.581%)	.008	.970	.226	-.0006 (0.594%)	.008	.964	.188	-.0023 (2.297%)	.007	.966

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B22

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.356	.0002 (-0.196%)	.004	.956	.370	-.0002 (0.226%)	.004	.955	.336	.0005 (-0.533%)	.004	.960
	Yes	.358	.0002 (-0.207%)	.004	.956	.372	-.0002 (0.215%)	.004	.955	.335	.0005 (-0.540%)	.004	.960
-0.10	No	.324	.0032 (-3.186%)	.004	.961	.400	.0003 (-0.268%)	.004	.964	.337	-.0022 (2.245%)	.004	.970
	Yes	.314	.0033 (-3.316%)	.004	.961	.387	.0003 (-0.300%)	.004	.964	.336	-.0021 (2.110%)	.004	.970
-0.15	No	.335	-.0002 (0.162%)	.004	.961	.385	-.0016 (1.575%)	.004	.957	.341	-.0018 (1.799%)	.004	.956
	Yes	.340	-.0002 (0.183%)	.004	.960	.387	-.0015 (1.537%)	.004	.957	.340	-.0018 (1.754%)	.004	.956

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B23

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.586	-.0018 (0.888%)	.009	.952	.642	-.0010 (0.487%)	.008	.958	.613	-.0024 (1.193%)	.008	.957
	Yes	.580	-.0018 (0.877%)	.009	.952	.641	-.0009 (0.442%)	.008	.959	.613	-.0023 (1.168%)	.008	.957
-0.10	No	.583	.0026 (-1.289%)	.009	.946	.617	.0049 (-2.471%)	.007	.965	.572	.0038 (-1.905%)	.007	.971
	Yes	.580	.0026 (-1.317%)	.009	.945	.629	.0050 (-2.518%)	.007	.965	.571	.0039 (-1.942%)	.007	.971
-0.15	No	.589	.0019 (-0.944%)	.008	.972	.625	.0014 (-0.688%)	.008	.955	.584	.0042 (-2.103%)	.008	.956
	Yes	.593	.0021 (-1.031%)	.008	.972	.629	.0015 (-0.744%)	.008	.955	.583	.0043 (-2.160%)	.008	.956

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B24

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
-0.05	No	.860	.0012 (-0.586%)	.004	.961	.909	.0011 (-0.531%)	.004	.971	.874	.0036 (-1.806%)	.004	.961
	Yes	.859	.0012 (-0.601%)	.004	.961	.902	.0011 (-0.537%)	.004	.971	.874	.0036 (-1.819%)	.004	.961
-0.10	No	.857	.0032 (-1.602%)	.005	.951	.887	-.0005 (0.271%)	.004	.962	.898	-.0021 (1.038%)	.004	.964
	Yes	.859	.0033 (-1.641%)	.005	.951	.890	-.0005 (0.241%)	.004	.962	.899	-.0020 (1.010%)	.004	.964
-0.15	No	.888	-.0010 (0.493%)	.004	.967	.894	.0012 (-0.613%)	.004	.967	.895	-.0041 (2.064%)	.004	.956
	Yes	.894	-.0008 (0.407%)	.004	.967	.894	.0013 (-0.671%)	.004	.967	.895	-.0040 (2.008%)	.004	.956

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B25

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.044	-.0104	.007	.957
	Yes	.062	-.0123	.008	.963	.069	-.0147	.008	.955	.048	-.0190	.007	.952
0.10	No	---	---	---	---	---	---	---	---	.039	-.0203	.007	.963
	Yes	.072	-.0381	.009	.945	.066	-.0343	.007	.950	.055	-.0375	.007	.947
0.15	No	---	---	---	---	---	---	---	---	.057	-.0290	.007	.967
	Yes	.092	-.0572	.008	.926	.119	-.0554	.008	.909	.096	-.0536	.007	.929

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B26

Type I Error Rate, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	α	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.042	-.0099	.004	.959
	Yes	.079	-.0190	.005	.940	.069	-.0186	.004	.947	.045	-.0183	.004	.956
0.10	No	---	---	---	---	---	---	---	---	.045	-.0210	.004	.955
	Yes	.088	-.0358	.004	.936	.100	-.0361	.004	.933	.075	-.0379	.004	.925
0.15	No	---	---	---	---	---	---	---	---	.099	-.0292	.004	.927
	Yes	.141	-.0509	.005	.882	.147	-.0579	.004	.881	.163	-.0540	.004	.871

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. α denotes Type I error rates, which were based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so $b(\hat{\phi})$ is the bias and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter.

Table B27

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.234	-.0130 (12.956%)	.008	.961
	Yes	.264	-.0183 (18.321%)	.008	.960	.285	-.0182 (18.229%)	.008	.957	.268	-.0204 (20.409%)	.008	.956
0.10	No	---	---	---	---	---	---	---	---	.251	-.0213 (21.258%)	.007	.969
	Yes	.301	-.0310 (31.045%)	.008	.959	.363	-.0375 (37.480%)	.008	.947	.331	-.0366 (36.604%)	.007	.953
0.15	No	---	---	---	---	---	---	---	---	.315	-.0277 (27.734%)	.007	.951
	Yes	.414*	-.0584 (58.400%)	.009	.921	.438*	-.0547 (54.733%)	.009	.915	.415*	-.0495 (49.505%)	.007	.937

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so

$b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter. * The power values cannot be appropriately interpreted due to the Type I error rate inflation.

Table B28

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.395	-.0113 (11.282%)	.004	.960
	Yes	.442*	-.0168 (16.82%)	.004	.952	.490	-.0171 (17.148%)	.004	.954	.447	-.0188 (18.824%)	.004	.959
0.10	No	---	---	---	---	---	---	---	---	.479	-.0252 (25.160%)	.004	.948
	Yes	.563*	-.0334 (33.427%)	.004	.936	.594*	-.0358 (35.803%)	.004	.926	.587	-.0403 (40.279%)	.004	.921
0.15	No	---	---	---	---	---	---	---	---	.570*	-.0310 (31.007%)	.004	.939
	Yes	.648*	-.0538 (53.811%)	.005	.877	.716*	-.0541 (54.050%)	.004	.887	.708*	-.0528 (52.788%)	.004	.894

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so

$b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter. * The power values cannot be appropriately interpreted due to the Type I error rate inflation.

Table B29

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.630	-.0111 (5.565%)	.008	.965
	Yes	.692	-.0199 (9.963%)	.009	.961	.705	-.0166 (8.322%)	.008	.958	.667	-.0180 (9.000%)	.008	.962
0.10	No	---	---	---	---	---	---	---	---	.674	-.0247 (12.341%)	.008	.958
	Yes	.691	-.0286 (14.315%)	.008	.956	.765	-.0335 (16.738%)	.008	.947	.738	-.0379 (18.946%)	.008	.947
0.15	No	---	---	---	---	---	---	---	---	.766	-.0356 (17.778%)	.008	.944
	Yes	.779*	-.0499 (24.958%)	.009	.935	.813*	-.0509 (25.433%)	.009	.922	.828*	-.0549 (27.438%)	.008	.923

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so

$b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter. * The power values cannot be appropriately interpreted due to the Type I error rate inflation.

Table B30

Power, Bias, Estimated Variance, and Coverage Rate of the Estimated Latent Mean Difference of the General Factor for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data				3-point scale data				5-point scale data			
		Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%	Power	$b(\hat{\phi})$	$\sigma^2_{\hat{\phi}}$	%
0.05	No	---	---	---	---	---	---	---	---	.921	-.0131 (6.527%)	.004	.954
	Yes	.917*	-.0183 (9.136%)	.005	.942	.935	-.0161 (8.066%)	.004	.954	.931	-.0198 (9.897%)	.004	.949
0.10	No	---	---	---	---	---	---	---	---	.944	-.0212 (10.577%)	.004	.955
	Yes	.951*	-.0350 (17.520%)	.004	.934	.969*	-.0328 (16.385%)	.004	.940	.963	-.0343 (17.144%)	.004	.935
0.15	No	---	---	---	---	---	---	---	---	.961*	-.0325 (16.235%)	.004	.929
	Yes	.971*	-.0546 (27.288%)	.004	.889	.987*	-.0537 (26.840%)	.004	.899	.985*	-.0514 (25.702%)	.004	.877

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. Power denotes empirical power based on significance level of .05. $\hat{\phi}$ denotes the estimated latent mean difference of the general factor, so

$b(\hat{\phi})$ is the bias (the values inside the parentheses denotes relative bias) and $\sigma^2_{\hat{\phi}}$ is the variance of this parameter. * The power values cannot be appropriately interpreted due to the Type I error rate inflation.

Table B31

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.999	.940/.969	.009/.008	.999/.999	.789/.842	.012/.010	.999/.999	.792/.863	.014/.009
	Yes	.998/.999	.997/1.024	.011/.010	.999/.999	.813/.864	.013/.010	.999/.999	.820/.889	.014/.010
0.10	No	.999/.999	.943/.972	.010/.009	.999/.999	.790/.842	.012/.009	.999/.999	.797/.864	.014/.009
	Yes	.997/.998	1.050/1.077	.017/.016	.999/.999	.827/.878	.014/.011	.998/.999	.846/.910	.017/.011
0.15	No	.999/.999	.947/.977	.010/.009	.999/.999	.796/.848	.012/.010	.998/.999	.806/.873	.015/.010
	Yes	.995/.995	1.131/1.157	.027/.025	.998/.999	.853/.902	.018/.013	.997/.999	.890/.953	.023/.015

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B32

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, $\Delta\kappa$ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.999	.950/.982	.008/.007	.999/.999	.810/.858	.012/.008	.999/1.000	.822/.886	.013/.008
	Yes	.999/.999	1.020/1.049	.010/.010	.999/.999	.836/.883	.012/.009	.999/.999	.854/.916	.014/.009
0.10	No	.999/.999	.952/.980	.008/.007	.999/.999	.810/.858	.011/.008	.999/.999	.828/.893	.014/.008
	Yes	.997/.998	1.117/1.141	.019/.017	.999/.999	.859/.905	.014/.010	.998/.999	.898/.959	.018/.011
0.15	No	.999/.999	.951/.980	.008/.007	.999/.999	.816/.863	.011/.008	.999/.999	.828/.887	.013/.008
	Yes	.995/.996	1.247/1.269	.027/.025	.998/.999	.903/.947	.018/.014	.997/.999	.964/1.015	.023/.015

Note: N denotes total sample size. *$\Delta\kappa$ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B33

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.998	.938/1.003	.009/.012	.999/.998	.788/.892	.012/.015	.999/.998	.790/.943	.014/.016
	Yes	.999/.998	.995/1.057	.011/.014	.999/.998	.812/.914	.012/.016	.999/.998	.818/.968	.014/.016
0.10	No	.999/.998	.944/1.006	.010/.013	.999/.998	.792/.899	.012/.016	.999/.998	.796/.945	.014/.016
	Yes	.997/.997	1.051/1.107	.018/.019	.999/.998	.830/.932	.015/.017	.998/.998	.844/.986	.017/.018
0.15	No	.999/.998	.944/1.004	.010/.013	.999/.998	.796/.896	.012/.015	.999/.998	.799/.935	.014/.015
	Yes	.995/.995	1.128/1.179	.026/.027	.998/.998	.853/.947	.017/.018	.997/.998	.881/1.006	.022/.019

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B34

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, $\Delta\kappa$ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.998	.954/1.050	.008/.013	.999/.998	.809/.967	.012/.017	.999/.998	.819/1.027	.013/.016
	Yes	.999/.998	1.023/1.114	.011/.015	.999/.998	.836/.990	.012/.017	.999/.998	.851/1.054	.014/.017
0.10	No	.999/.998	.952/1.047	.008/.013	.999/.998	.811/.964	.011/.016	.999/.998	.826/1.052	.013/.018
	Yes	.997/.997	1.120/1.202	.019/.021	.999/.998	.863/1.009	.014/.018	.998/.998	.897/1.109	.018/.020
0.15	No	.999/.998	.952/1.044	.008/.012	.999/.998	.816/.966	.011/.016	.999/.998	.830/1.035	.014/.017
	Yes	.995/.994	1.255/1.327	.028/.029	.998/.998	.901/1.039	.018/.020	.997/.998	.963/1.146	.023/.022

Note: N denotes total sample size. *$\Delta\kappa$ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B35

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.996	.941/1.094	.010/.023	.999/.995	.789/1.048	.012/.031	.999/.995	.788/1.157	.013/.032
	Yes	.998/.996	.998/1.144	.012/.023	.999/.995	.814/1.067	.012/.030	.999/.995	.817/1.177	.014/.032
0.10	No	.999/.996	.943/1.098	.010/.023	.999/.995	.787/1.042	.011/.030	.999/.995	.796/1.138	.014/.031
	Yes	.997/.994	1.048/1.190	.017/.028	.999/.995	.824/1.070	.013/.031	.998/.995	.845/1.173	.017/.032
0.15	No	.999/.996	.947/1.094	.010/.023	.999/.995	.797/1.040	.012/.030	.999/.995	.799/1.140	.013/.031
	Yes	.995/.992	1.124/1.252	.026/.033	.998/.994	.854/1.084	.018/.032	.997/.994	.880/1.198	.022/.034

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B36

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in General Factor Loadings (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.996	.955/1.219	.008/.025	.999/.995	.810/1.243	.011/.033	.999/.995	.824/1.405	.013/.034
	Yes	.999/.995	1.029/1.278	.011/.026	.999/.995	.837/1.261	.012/.033	.999/.995	.857/1.425	.014/.034
0.10	No	.999/.996	.956/1.220	.008/.025	.999/.995	.812/1.227	.011/.032	.999/.995	.824/1.409	.013/.034
	Yes	.997/.994	1.119/1.353	.019/.030	.999/.995	.861/1.259	.014/.033	.998/.994	.894/1.450	.018/.035
0.15	No	.999/.996	.956/1.203	.008/.024	.999/.995	.817/1.218	.011/.032	.999/.995	.833/1.392	.014/.033
	Yes	.995/.992	1.257/1.456	.028/.035	.998/.994	.906/1.277	.019/.034	.997/.994	.967/1.474	.023/.036

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B37

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.999	.965/.991	.010/.009	.999/.999	.787/.837	.012/.009	.999/.999	.798/.870	.014/.009
	Yes	.999/.999	.986/1.012	.010/.009	.999/.999	.808/.857	.012/.009	.999/.999	.815/.886	.014/.009
0.10	No	.999/.999	.964/.995	.010/.009	.999/.999	.789/.840	.012/.009	.999/.999	.800/.872	.014/.009
	Yes	.999/.999	.995/1.024	.011/.011	.999/.999	.819/.869	.014/.010	.998/.999	.829/.899	.016/.010
0.15	No	.999/.999	.968/.999	.010/.009	.999/.999	.789/.839	.012/.009	.999/.999	.795/.863	.013/.008
	Yes	.998/.998	1.015/1.045	.013/.012	.998/.999	.836/.883	.016/.012	.998/.999	.845/.909	.018/.011

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B38

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, $\Delta\kappa$ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.999	.970/1.001	.008/.007	.999/.999	.810/.863	.012/.009	.999/.999	.826/.894	.013/.008
	Yes	.999/.999	.994/1.024	.008/.008	.999/.999	.835/.887	.012/.009	.999/.999	.846/.913	.014/.009
0.10	No	.999/.999	.967/.996	.007/.007	.999/.999	.809/.861	.011/.008	.999/1.000	.822/.888	.012/.008
	Yes	.999/.999	1.010/1.038	.010/.009	.999/.999	.852/.901	.014/.010	.999/.999	.864/.928	.016/.009
0.15	No	.999/.999	.970/1.001	.007/.007	.999/1.000	.806/.854	.011/.008	.999/.999	.827/.896	.013/.008
	Yes	.999/.999	1.046/1.075	.012/.011	.998/.999	.881/.926	.017/.012	.998/.999	.905/.969	.019/.012

Note: N denotes total sample size. *$\Delta\kappa$ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B39

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.998	.964/1.032	.010/.013	.999/.998	.789/.897	.012/.015	.999/.998	.795/.947	.013/.016
	Yes	.999/.998	.984/1.051	.010/.014	.999/.998	.810/.916	.013/.016	.999/.998	.812/.962	.013/.016
0.10	No	.999/.998	.966/1.025	.010/.013	.999/.998	.786/.893	.011/.015	.999/.998	.798/.952	.013/.016
	Yes	.998/.998	.997/1.055	.011/.014	.999/.998	.818/.921	.013/.016	.998/.998	.828/.977	.016/.017
0.15	No	.999/.998	.964/1.021	.010/.012	.999/.998	.791/.900	.012/.016	.999/.998	.799/.951	.013/.016
	Yes	.998/.998	1.010/1.065	.013/.015	.998/.998	.839/.944	.016/.018	.998/.998	.848/.993	.019/.018

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B40

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.999	.965/1.061	.007/.012	.999/.998	.808/.971	.011/.017	.999/.998	.828/1.050	.013/.017
	Yes	.999/.999	.988/1.083	.008/.012	.999/.998	.831/.991	.012/.017	.999/.998	.848/1.066	.014/.017
0.10	No	.999/.999	.969/1.061	.008/.012	.999/.998	.809/.972	.011/.017	.999/.998	.825/1.053	.013/.017
	Yes	.999/.998	1.012/1.100	.010/.013	.999/.998	.851/1.008	.014/.018	.999/.998	.866/1.086	.016/.019
0.15	No	.999/.999	.970/1.065	.008/.012	.999/.998	.809/.978	.011/.017	.999/.998	.823/1.051	.012/.017
	Yes	.999/.998	1.045/1.134	.012/.016	.998/.998	.882/1.040	.017/.020	.998/.998	.902/1.116	.019/.020

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B41

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.996	.961/1.124	.009/.023	.999/.995	.787/1.057	.012/.031	.999/.995	.795/1.168	.013/.032
	Yes	.999/.996	.981/1.141	.010/.023	.999/.995	.809/1.074	.012/.031	.999/.995	.813/1.181	.014/.032
0.10	No	.999/.996	.961/1.120	.009/.023	.999/.995	.786/1.042	.012/.030	.999/.995	.797/1.151	.013/.031
	Yes	.999/.995	.992/1.147	.011/.024	.999/.995	.817/1.066	.013/.031	.999/.995	.825/1.171	.015/.032
0.15	No	.999/.996	.962/1.119	.009/.023	.999/.995	.792/1.056	.012/.031	.999/.995	.800/1.157	.013/.031
	Yes	.998/.995	1.009/1.160	.012/.025	.998/.994	.840/1.093	.017/.033	.998/.994	.849/1.192	.019/.033

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B42

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Specific Factor Loadings (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	.999/.996	.967/1.240	.008/.025	.999/.995	.806/1.248	.011/.033	.999/.995	.825/1.412	.013/.034
	Yes	.999/.996	.991/1.258	.008/.024	.999/.995	.831/1.265	.012/.033	.999/.995	.845/1.424	.014/.034
0.10	No	.999/.996	.972/1.241	.008/.025	.999/.995	.805/1.255	.011/.034	.999/.995	.823/1.435	.013/.035
	Yes	.999/.995	1.015/1.275	.010/.025	.999/.995	.846/1.282	.013/.034	.999/.994	.864/1.459	.015/.035
0.15	No	.999/.996	.970/1.246	.007/.025	.999/.995	.808/1.250	.011/.033	.999/.994	.819/1.442	.012/.035
	Yes	.999/.995	1.042/1.303	.012/.027	.998/.994	.884/1.301	.017/.035	.998/.994	.899/1.491	.019/.036

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B43

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.999	.785/.845	.014/.009
	Yes	.999/.999	.982/1.012	.010/.009	.999/.999	.809/.862	.013/.010	.999/.999	.817/.886	.015/.009
0.10	No	---	---	---	---	---	---	.998/.999	.799/.862	.016/.011
	Yes	.999/.999	.990/1.024	.011/.011	.998/.999	.834/.885	.017/.013	.998/.999	.846/.922	.019/.013
0.15	No	---	---	---	---	---	---	.998/.999	.826/.888	.020/.013
	Yes	.998/.998	1.007/1.045	.013/.013	.998/.998	.863/.929	.021/.018	.997/.998	.890/.972	.025/.017

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B44

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = 0)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.999	.817/.881	.014/.009
	Yes	.999/.999	.990/1.024	.008/.008	.999/.999	.835/.886	.013/.009	.999/.999	.855/.930	.015/.010
0.10	No	---	---	---	---	---	---	.999/.999	.844/.909	.017/.011
	Yes	.999/.999	1.012/1.048	.010/.010	.999/.999	.875/.937	.017/.013	.998/.999	.908/.994	.020/.014
0.15	No	---	---	---	---	---	---	.998/.999	.896/.967	.022/.015
	Yes	.999/.999	1.044/1.091	.013/.013	.998/.998	.933/1.011	.023/.019	.997/.998	.993/1.097	.027/.020

Note: N denotes total sample size. *Δκ = 0* denotes that the latent mean difference of the general factor was generated to be 0. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B45

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.998	.790/.946	.015/.018
	Yes	.999/.998	.980/1.055	.010/.014	.999/.998	.815/.944	.014/.019	.999/.998	.822/1.007	.016/.020
0.10	No	---	---	---	---	---	---	.999/.998	.798/.959	.016/.020
	Yes	.999/.998	.990/1.074	.011/.016	.999/.997	.826/.980	.015/.023	.998/.997	.844/1.049	.019/.023
0.15	No	---	---	---	---	---	---	.998/.997	.826/.993	.021/.023
	Yes	.998/.996	1.009/1.117	.013/.021	.998/.996	.860/1.038	.021/.028	.997/.996	.889/1.113	.025/.028

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B46

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = -0.10)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.998	.821/1.050	.014/.020
	Yes	.999/.998	.993/1.111	.008/.015	.999/.998	.837/1.034	.012/.020	.999/.998	.860/1.136	.015/.021
0.10	No	---	---	---	---	---	---	.999/.998	.844/1.108	.017/.022
	Yes	.999/.998	1.011/1.150	.010/.018	.999/.997	.873/1.110	.017/.025	.998/.997	.906/1.244	.020/.026
0.15	No	---	---	---	---	---	---	.998/.997	.894/1.162	.022/.026
	Yes	.999/.997	1.044/1.214	.013/.022	.998/.996	.935/1.207	.023/.030	.997/.996	.989/1.349	.027/.031

Note: N denotes total sample size. *Δκ = -0.10* denotes that the latent mean difference of the general factor was generated to be -0.10. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B47

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=600, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.995	.786/1.144	.014/.034
	Yes	.999/.995	.983/1.165	.010/.026	.999/.994	.813/1.107	.013/.034	.999/.994	.818/1.228	.015/.035
0.10	No	---	---	---	---	---	---	.998/.994	.803/1.189	.017/.037
	Yes	.998/.994	.996/1.187	.012/.028	.998/.993	.833/1.156	.016/.038	.998/.993	.848/1.304	.019/.040
0.15	No	---	---	---	---	---	---	.998/.993	.825/1.229	.020/.040
	Yes	.998/.993	1.010/1.229	.013/.032	.998/.992	.863/1.215	.021/.043	.997/.991	.889/1.379	.025/.045

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.

Table B48

Goodness of Fit Indices for the Generated Multiple-group Bifactor IRT Models with DIF in Threshold Parameters (N=1200, Δκ = -0.20)

Magnitude of DIF	Equality Constraints on Parameters with DIF	Binary data			3-point scale data			5-point scale data		
		CFI	WRMR	RMSEA	CFI	WRMR	RMSEA	CFI	WRMR	RMSEA
0.05	No	---	---	---	---	---	---	.999/.995	.818/1.414	.014/.036
	Yes	.999/.995	.993/1.306	.008/.027	.999/.994	.843/1.333	.013/.037	.999/.994	.856/1.534	.015/.038
0.10	No	---	---	---	---	---	---	.999/.994	.849/1.462	.017/.038
	Yes	.999/.994	1.014/1.360	.010/.030	.999/.993	.878/1.411	.017/.040	.998/.993	.910/1.630	.020/.041
0.15	No	---	---	---	---	---	---	.998/.993	.897/1.534	.022/.041
	Yes	.999/.993	1.048/1.432	.013/.034	.998/.991	.934/1.521	.023/.045	.997/.991	.989/1.757	.027/.046

Note: N denotes total sample size. *Δκ = -0.20* denotes that the latent mean difference of the general factor was generated to be -0.20. For each condition, the first value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was freely estimated, and the second value denotes the goodness of fit index for the model in which the latent mean difference of the general factor was constrained to be 0.