TESTING MEASUREMENT INVARIANCE ON ZERO-INFLATED MEASURES WITH TWO-PART FACTOR MODEL AND MGCFA

A Dissertation

by

MIRIM KIM

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Myeongsun Yoon
Committee Members,	Michael Beyerlein
	Robert Hall
	Oi-Man Kwok
	Wen Luo
Head of Department,	Fuhui Tong

August 2021

Major Subject: Educational Psychology

ABSTRACT

Measurement invariance justifies using observed variables to identify group differences on the latent construct because it indicates that there is no systematic measurement bias between the observed and the latent variables across groups. Using structural equation modeling, multi-group confirmatory factor analysis (MGCFA) is a commonly used tool for testing measurement invariance. However, to date, the usefulness of MGCFA for testing measurement invariance on zero-inflated variables has not been studied. Due to the non-normality of zero-inflated data, flexible modeling to handle zero-inflation and extending the two-part model to factor analysis (two-part factor model) is possible. Therefore, we examined how different levels of zero-inflation affected the measurement invariance tests with the two-part factor model and the MGCFA and suggested the appropriate factor analysis model when zero-inflated variables are the target measures.

Study I compared the performance of the two-part factor model and the MGCFA on testing measurement invariance of empirical zero-inflated data. The two models led to different measurement invariance results on the target variables. Thus, applying a different factor model to test measurement invariance brought different conclusions when the measures were zero-inflated. Study II evaluated the performances of the two-part factor model and the MGCFA across different simulation conditions: sample size, level of non-invariance, and extent of zero-inflation. Both models showed acceptable Type I error rates except for some conditions, and the two-part factor model outperformed MGCFA in terms of its power to detect non-invariance on the zero-inflated variable. However, the two models had low power and difficulty in identifying correct partial invariance models when the zero-inflation was extreme (90%).

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by dissertation committees consisting of Dr. Myeongsun Yoon, Dr. Oi-Man Kwok, Dr. Wen Luo, and Dr. Robert Hall of the Department of Educational Psychology and Dr. Michael Beyerlein of the Department of Educational Administration and Human Resource Development.

All work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the College of Education and Human Development.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

Comparing groups of outcomes is a common approach to deriving meaningful conclusions from a target variable. If the target variable is an observable attribute, such as weight or height in the natural sciences, different results across groups directly indicate true group differences. By contrast, social sciences research frequently discusses unobservable and latent attributes (e.g., motivation or achievement) across groups, yet uses group comparisons of the observed scores to represent latent scores differences. However, to justify using observed scores to indicate latent scores, the relationship between the observed and the latent scores must be systematically derived without measurement bias. That is, the latent variable must be measured identically across different groups to ensure the observed score is the same if individuals have identical latent scores but are in different groups (Horn & McArdle, 1992; Millsap, 2012; Sörbom, 1974). This conditional independence of the observed variables given the latent variables, irrespective of group membership, is referred to as measurement invariance (Kim & Yoon, 2011b). In contrast, if measurement process differs depending on the groups, and the measurement invariance assumption is not met, the observed score of individuals with the same latent score but different group membership may differ. In other words, the observed difference score between groups might not represent an actual group difference on the latent score.

Various statistical approaches are used to test measurement invariance. Millsap (2012) classified these approaches as (a) observed variable methods, such as Mantel-Haenszel method or SIBTEST (Shealy & Stout, 1993); and (b) latent variable models using structural equation modeling (SEM) or an item response theory framework (Kim & Yoon, 2011a; Lord & Novick, 2008; Millsap, 2012; Stark, Chernyshenko, & Drasgow, 2006; Woods, 2009).

Among these, multi-group confirmatory factor analysis (MGCFA), a typical SEM model, is widely used. Several issues related to MGCFA have been investigated in the literature, leading to recommendations for the most appropriate approach. For example, types of measurement (i.e., categorical measure or continuous measure) (Kim & Yoon, 2011a; Liu et al., 2016; Millsap & Yun-Tein, 2004) and complex data structures, such as multilevel data (Kim, Kwok, & Yoon, 2012; Ryu, 2014) and longitudinal data (Kim & Willson, 2014; Liu et al., 2016) have been studied. Also, partial invariance (Jung & Yoon, 2016, 2017; van de Schoot et al., 2013; Yoon & Millsap, 2007) and fit indices for testing measurement invariance (Cheung & Rensvold, 2002; Fan & Sivo, 2009; Lai & Yoon, 2014; Meade & Bauer, 2007) have been examined. The diverse topics related to measurement invariance imply the importance of measurement invariance.

Despite the existing research, it is hard to find guidelines for testing measurement invariance dealing with non-normal data. Furthermore, limited research has focused on testing measurement invariance on zero-inflated data, a special type of non-normal data. As a result, it is unclear whether typical measurement invariance tests work properly with non-normal, especially zero-inflated, data. Several studies have proposed using robust estimators to handle non-normality in measurement invariance tests, but they have mainly involved non-normality from ordinal-categorical measurement (Kim & Yoon, 2011a; Liu et al., 2016), not inflated item responses. Further, while many studies have discussed strategies for handling non-normality, they are not closely tied to measurement invariance tests. As a general methodological strategy for non-normality, data transformation (Howell, 2012; Yuan, Chan, & Bentler, 2000) or robust estimator is well known for achieving less biased estimates (Curran, West, & Finch, 1996; Satorra, 1990; Savalei, 2014; Yuan & Bentler, 2000). A more flexible model represents yet another way to explain non-normal data distribution. Further studies are required to implement such strategies to test the measurement invariance on the zero-inflated variables.

Zero-inflation occurs due to excessive responses on the zero value because most people express a lack of experience with zero. For example, if items ask about rare experiences (e.g., bullying/victimization experiences for the past year), item responses are mostly 0; therefore, the item distribution will be skewed/inflated to zero. Although such zero-inflated items measure unusual events, that does not necessarily mean that the related research is also uncommon. In fact, zero-inflation is a common issue in delinquency (Brown, Catalano, Fleming, Haggerty, & Abbott, 2005; Ferrer, Conger, & Robins, 2016; Kaysen et al., 2014; Kim & Muthén, 2009; Liu, Ma, & Johnson, 2008) and public health or health economics (Deb & Holmes, 2000; Deb & Trivedi, 2002; Duan, Manning, Morris, & Newhouse, 1983; Liu, Strawderman, Cowen, & Shih, 2010). Comparing groups of such zero-inflated variables is also common, yet measurement invariance has not been discussed enough. To fill this gap in the literature, the two issues of measurement invariance and zero-inflation are taken together in our study. That is, when examining the measurement invariance of zero-inflated data, we look at which flexible model is more appropriate for handling zero-inflation. Starting from this point, the current study suggests a useful factor analysis model for testing measurement invariance when normality is not met due to zero-inflation.

One of the best known models for handling zero-inflation is the two-part model. The core idea of two-part modeling is to decompose the majority of zeros and the minority of non-zeros into two parts to lessen the zero-inflation (Duan, Willard G. Manning, Morris, & Newhouse, 1984; Kim & Muthén, 2009; Xu, Paterson, Turpin, & Xu, 2015). We focus on this model because implementing the two-part modeling in the SEM is possible (Brown et al., 2005; Ferrer et al., 2016; Kim & Muthén, 2009; McTernan & Blozis, 2014; Muthén, 2001). However, to date, there is a lack of in-depth research on the performance of two-part modeling in terms of testing measurement invariance.

Previously, Kim and Muthén (2009) demonstrated the use of two-part factor mixture modeling on testing measurement invariance. However, their study did not evaluate the performance of two-part factor mixture modeling and the possible factors influencing this model. Therefore, the present study investigates whether or not the two-part modeling is appropriate for testing measurement invariance of zero-inflated variables under various data conditions. Furthermore, we compare the two-part factor model with the MGCFA and examine the effect of different extents of zero-inflation on the two models' performances. The following chapter reviews the literature on measurement invariance and two-part modeling as well as presents the purposes of the study.

CHAPTER II

LITERATURE REVIEW

This chapter overviews two primary topics: measurement invariance and methodological issues related to zero-inflated data and research purposes. Specifically, the following issues are reviewed: (1) definition of measurement invariance; (2) factorial invariance; (3) methods to detect measurement invariance; (4) general models to handle zero-inflated data; and (5) two-part factor modeling as a tool for testing measurement invariance on zero-inflated data.

Definition of Measurement Invariance

Researchers use observed variables (i.e., measures or items) to discuss latent constructs, such as traits or attributes, by assuming the observed score represents the level of underlying latent variable. To justify the uses of the observed variables, a systematically accurate measurement process of assigning numbers to the observed variable is required to reflect the corresponding level of the latent variable. In other words, the measurement needs to build the relationship between an observed score and a latent score without measurement bias (Millsap, 2012; Vandenberg & Lance, 2000).

We look at the measurement bias coming from an individual's group membership in this study; therefore, the present study focuses on whether or not the observed score is from an equal function between the observed and the latent variables regardless of individuals' groups. Suppose there is a systematic difference in the measurement according to groups (e.g., occasion, sex, ethnicity, age, culture, and so on). In that case, the observed score for the same level of the latent variable may differ due to the different subgroups; therefore, the measurement would be biased in terms of the groups. This measurement bias is referred to as measurement non-invariance. In contrast, when the observed score is identical given an equivalent level of latent construct regardless of the subgroups, it indicates measurement invariance (Meredith & Millsap, 1992; Millsap, 2012; Vandenberg & Lance, 2000).

Measurement invariance supports the measurement quality of developed measures and justifies using measured variables for group comparisons because there is no measurement bias from the groups. On the other hand, if measurement invariance does not hold, the measures have inadequate validity (Millsap, 2012; Vandenberg & Lance, 2000), and discussing group differences on the target latent construct may mislead the study. Without measurement invariance, it is not clear that whether an observed difference score is from different latent scores across groups or the systematically inaccurate measurement related to the group memberships (Cheung & Rensvold, 2002; Horn & McArdle, 1992; Little, 1997; Stark et al., 2006).

Factorial Invariance

As a special type of measurement invariance, factorial invariance has to do with measured variables an underlying factor structure (Millsap, 2012). Testing factorial invariance with a factor analysis model is common, and confirmatory factor analysis (CFA) model is widely used. The factor model is a linear regression model where the observed measures are regressed on the latent factors as follows:

$$X = \tau + \Lambda\xi + \delta \tag{1}$$

X represents a $p \times 1$ vector of measures, τ and δ are $p \times 1$ vectors of intercepts and unique factors respectively. In terms of the regression model, τ are the regression intercepts and δ are the regression residuals. Λ indicates a $p \times m$ matrix for factor loadings or regression slopes, and ξ is a $m \times 1$ vector of factor scores. Also, X is factorial when the factor analysis model holds (Meredith, 1993), and the mean vector (μ) and covariance matrix (Σ) of X are as follows:

$$E(X) = \mu = \tau + \Lambda\kappa \quad (2)$$

$$Cov(X) = \Sigma = \Lambda\Phi\Lambda' + \Theta \quad (3)$$

Equation 2 represents a $p \times 1$ mean vector of measures with κ denoting factor means, which is a mean vector of ξ. Equation 3 shows a $p \times p$ variance-covariance matrix of measures. Φ and θ indicate $p \times p$ variance-covariance matrix of ξ and δ respectively.

The above factor analysis model can be extended to multiple subgroups, and testing factorial invariance across subgroups is possible with the MGCFA (Meredith, 1993; Millsap, 2012). All parameters in Equations 1, 2, and 3 are allowed to be different depending on the group membership (g). Equations of MGCFA for gth subgroup are expressed as follows:

$$X_g = \tau_g + \Lambda_g\xi_g + \delta_g \quad (4)$$

$$\mu_g = \tau_g + \Lambda_g\kappa_g \quad (5)$$

$$\Sigma_g = \Lambda_g\Phi_g\Lambda_g' + \Theta_g \quad (6)$$

If factorial invariance holds, then specific parameters (Λ_g, τ_g, and Θ_g) representing a factorial relationship are invariant across groups. Testing factorial invariance is sequential as placing the gradual invariant restrictions on the parameters. Depending on the restriction, the type of factorial invariance model varies.

A Sequence of Models for Factorial Invariance

Vandenberg and Lance (2000) defined the sequence of the factorial invariance test. Commonly, four models: configural, metric, scalar, and strict invariance, are sequentially tested with gradually constrained parameters across groups.

Configural Invariance

Configural invariance concerns whether the same factor structure is applicable to multiple groups by having identical patterns of factor loadings under the CFA model. Holding configural invariance indicates the CFA model examines the same number of factors, and corresponding measures are identically loaded on the specific factor regardless of subgroups. Configural invariance is necessary for a multi-group comparison because subgroups must have an equivalent model to be tested. The following hypothesis represent testing configural invariance,

$$H_{configural}: \Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Theta_g,\ \mu_g = \tau_g + \Lambda_g \kappa_g \tag{7}$$

Assume a test has six measured variables for two latent constructs of anxiety and depression, and each of the three variables measures one of the latent constructs. For the factorial invariance test of the six measures across two groups (i.e., sex: male and female), the parameters in Equation 7 can be shown as follows:

$$\Lambda_g: \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{211} & 0 \\ \lambda_{311} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{521} \\ 0 & \lambda_{621} \end{bmatrix}_{g1} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{212} & 0 \\ \lambda_{312} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{522} \\ 0 & \lambda_{622} \end{bmatrix}_{g2} \tag{8}$$

$$\Phi_g: \begin{bmatrix} \phi_{111} & \phi_{121} \\ \phi_{211} & \phi_{221} \end{bmatrix}_{g1} = \begin{bmatrix} \phi_{112} & \phi_{122} \\ \phi_{212} & \phi_{222} \end{bmatrix}_{g2} \tag{9}$$

$$\Theta_g: \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{31} \\ \theta_{41} \\ \theta_{51} \\ \theta_{61} \end{bmatrix}_{g1} = \begin{bmatrix} \theta_{12} \\ \theta_{22} \\ \theta_{32} \\ \theta_{42} \\ \theta_{52} \\ \theta_{62} \end{bmatrix}_{g2} \tag{10}$$

$$\tau_g: \begin{bmatrix} \tau_{11} \\ \tau_{21} \\ \tau_{31} \\ \tau_{41} \\ \tau_{51} \\ \tau_{61} \end{bmatrix}_{g1} = \begin{bmatrix} \tau_{12} \\ \tau_{22} \\ \tau_{32} \\ \tau_{42} \\ \tau_{52} \\ \tau_{62} \end{bmatrix}_{g2} \quad (11)$$

$$\kappa_g: \begin{bmatrix} \kappa_{11} \\ \kappa_{21} \end{bmatrix}_{g1} = \begin{bmatrix} \kappa_{12} \\ \kappa_{22} \end{bmatrix}_{g2} \quad (12)$$

With three subscripts (i.e., p, m, and g for measures, factors, and subgroups), parameters are differently estimated depending on the corresponding subscript. Elements of factor loading and factor variance-covariance matrix are represented as Λ_{pmg} and Φ_{ppg} respectively. Also, elements of the unique factor variance-covariance matrix are expressed as Θ_{pg} as Equation 10 if there is no correlation between unique factors. The τ_{pg} and κ_{mg} indicate the elements of intercept and factor mean vector respectively.

Under configural invariance, a reference variable scaling the latent factor has an invariant factor loading (λ_{pm}) across groups, as Equation 8 shows, and patterns of zero or non-zero factor loadings are the same (Horn & McArdle, 1992; Steenkamp & Baumgartner, 1998). However, there are no equal constraints on parameters across groups except for the reference variable; therefore, parameters have a different g subscript for each group. Factor loading locations need to be equivalent for all subgroups, and it shows measures that are related to a particular factor.

Metric Invariance

When configural invariance holds, testing metric invariance is feasible. The metric invariance model tests equal metrics across subgroups (Steenkamp & Baumgartner, 1998) with the following hypothesis,

$$H_{metric}\colon \Sigma_g = \Lambda \Phi_g \Lambda' + \Theta_g, \mu_g = \tau_g + \Lambda \kappa_g \tag{13}$$

$$\Lambda_g\colon \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}_{g1} = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix}_{g2} \tag{14}$$

Equation 14 represents the difference with configural invariance. Factor loadings are invariant for two groups; therefore, metric invariance indicates that any change in the measure is due to a change in the unit of the corresponding factor, and it is the same for groups (Lubke & Muthen, 2005). With the above example, each linear slope of the regression model between the measure and the factor is identical for both males and females. Metric invariance is also known as weak invariance (Meredith, 1993).

Scalar Invariance

The configural and metric invariance models concern invariant covariation of the factorial across groups. Therefore, metric invariance still does not validate that the observed mean difference score between groups is from the actual factor mean difference. That is because metric invariance does not test intercepts in the mean structure (μ_g). When group comparison is the main research goal, the study needs to take the invariant mean structure μ_g into account, and testing intercepts in the factor model is necessary in such cases (Geiser, 2013; Steenkamp & Baumgartner, 1998). Testing intercept invariance is followed once metric invariance holds, and the equal intercept constraints are as follows:

$$H_{scalar}\colon \Sigma_g = \Lambda \Phi_g \Lambda' + \Theta_g, \mu_g = \tau + \Lambda \kappa_g \tag{15}$$

$$\tau_g: \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix}_{g1} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \end{bmatrix}_{g2} \quad (16)$$

If intercepts are not invariant across groups, one of the subgroups has systematically higher or lower means of measures even though all subgroups have the same factor mean score. Consequently, it is difficult to say the observed mean differences among groups are utterly from the factor mean differences (Lubke & Muthen, 2005; Steenkamp & Baumgartner, 1998). In the example above, when scalar invariance holds, the observed difference scores between males and females are justified by the different factor mean scores. For the scalar invariance, Equations 14 and 16 hold simultaneously. Strong invariance (Meredith, 1993) is another term for scalar invariance.

Strict Invariance

Strict invariance has additional constraints than scalar invariance in terms of unique factor variances as follows:

$$H_{strict}: \Sigma_g = \Lambda \Phi_g \Lambda' + \Theta, \mu_g = \tau + \Lambda \kappa_g \quad (17)$$

$$\Theta_g: \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix}_{g1} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \end{bmatrix}_{g2} \quad (18)$$

The equal unique factor variance is similar to homoscedasticity (i.e., homogeneity of variance) in regression analysis (Meredith, 1993; Vandenberg & Lance, 2000). Homoscedasticity assumes the constant residual variance in the population. The violation of homogeneity of residual variances across samples is related to biased parameter estimates; therefore, it is required for the validity of the prediction (Cohen & Cohen, 1983). Likewise, strict invariance is examined for the validity of the measurement instrument. If Equations 14, 16, and 18 hold together, strict invariance holds.

Model Evaluation to Test Measurement Invariance

The evaluation of the factorial invariance test is based on fit indices by fitting sequentially constrained MGCFA models. In other words, a more restricted model is compared with a less restricted model based on the fit indices to see whether the new constraint on the previous model is statistically plausible or not. There are two types of fit indices to consider. First, global fit indices test an overall model with parameters that are equally constrained across groups. Second, local test statistics are taken into account for model comparisons when testing a specific parameter rather than overall model tests.

Global Fit Indices for Full Invariance

A full factorial invariance model refers to constraining all parameters to be equal across groups (Millsap, 2012; Steenkamp & Baumgartner, 1998; Yoon & Millsap, 2007). For example, if all factor loadings are invariant for all groups, it is full metric invariance. Testing full invariance is based on global fit indices.

A χ^2 test is the most common way to test the hypothesized model. In the SEM framework, the χ^2 test is to test the difference between a hypothesized model and an observed data. A model implied variance-covariance matrix from the hypothesized model, and the observed variance-covariance matrix from the data is compared, and the χ^2 tests a null hypothesis that there is no difference between the two matrices. The χ^2 statistic is 0 if the model is perfectly fit to the data. The model fit is getting worse when the tested model less adequately fits to the data. That is, χ^2

statistic is getting larger than 0. It is statistically significant to reject the hypothesized model, if the χ^2 statistic is greater than the cutoff χ^2 with the corresponding degree of freedom.

Although the hypothesized model is fitted to the data well, the χ^2 test is likely to reject the null model with large sample size, resulting in Type I error. Therefore, alternative fit indices less impacted by sample size are considered together. Comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root-mean-square residual (SRMR) are the commonly considered fit indices for SEM. Hu and Bentler (1999) investigated these fit indices and suggested practical cutoff criteria. According to Hu and Bentler (1999), the adequate cutoff values of CFI (Bentler, 1990; Lai & Yoon, 2014), RMSEA (Steiger, 1989, 2009), and SRMR (Bentler, 1995) are 0.95, 0.08, and 0.06 respectively. For example, a specified invariance model holds if CFI is above 0.95, RMSEA is below 0.08, and SRMR is below 0.06.

Test Statistics for Partial Invariance

Full invariance does not always happen in reality (Cheung & Lau, 2012; Vandenberg & Lance, 2000; Yoon & Millsap, 2007). In that case, the study focuses on the part of parameter invariance rather than full invariance. Factorial invariance is referred to as partial factorial invariance when some of the parameters are constrained to be equal across groups (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). In other words, partial invariance indicates partially invariant factor loadings, intercepts, or unique factor variances.

The baseline model determines a direction of testing partial invariance, and the partial equivalence can be either gradually freed or constrained depending on the direction. Suppose a baseline model tests full factorial invariance (i.e., full metric, scalar, or strict invariance) but fails to hold it. In that case, the alternative model aims to test partial factorial invariance with fewer constraints on the parameters than the baseline, and the full invariance model is nested within the partial invariance model. For example, the partial metric invariance model examines a freed factor loading of each variable, except for a reference variable, instead of constraining all factor loadings together if full metric invariance fails to hold. The model evaluation is based on the test statistic (e.g., likelihood ratio test) by comparing the partial metric invariance model to the baseline model (i.e., the full metric invariance model). The freely estimated factor loading is considered non-invariant if the test statistic is significant. When partial factorial invariance is examined based on the full invariance model, as a baseline model, and the parameters are gradually freed, this directional procedure is called a backward procedure (Millsap, 2012; Yoon & Millsap, 2007). The backward procedure compares less restricted partial factorial invariance models to the baseline model.

On the other hand, a baseline model can have all parameters freely estimated across groups, except for a reference variable. In this case, partial invariance models are more restricted than the baseline model; therefore, the partial model is nested within the baseline model. Comparing more restricted partial factorial invariance models to the baseline model that does not have any equal constraints on the parameters is called a forward procedure (Jung & Yoon, 2016; Kim, Joo, Lee, Wang, & Stark, 2016). If the test statistic is significant, the tested parameter is considered non-invariant. Both backward and forward procedures require repeated model comparisons; therefore, Bonferroni correction is used to control Type I error inflation (Millsap, 2012; Stark et al., 2006).

As test statistics for the model comparison between the nested models, likelihood ratio test (LRT), Wald test, and modification index (MI) are commonly used in SEM (Bollen, 1989; Chou & Bentler, 1990). Tests are based on a difference of degree of freedom (Δdf) to see either constrained or released parameters are statistically significant (Bollen, 1989; Chou & Bentler, 1990; Vandaele, 1981). The three statistics are also widely used for testing partial factorial invariance (Cheung & Lau, 2012; Kim et al., 2016; Meade & Bauer, 2007). In general, a single parameter is examined one at a time; that is, the test statistics evaluate the partial factorial model based on 1 degree of freedom (Δdf = 1).

A different baseline model prefers a different test statistic (Byrne et al., 1989; Chou & Bentler, 1990; Jung & Yoon, 2017; Kim et al., 2016; Yoon & Millsap, 2007). The LRT is used in both forward and backward procedures; therefore, the baseline model can be either the full factorial invariance model or the non-invariance model (Byrne et al., 1989; Kim et al., 2016). On the other hand, the Wald test is used in the forward procedure (Jung & Yoon, 2017), and the MI is used for the backward procedure(Yoon & Millsap, 2007).

Likelihood Ratio Test (LRT)

The LRT is based on a loglikelihood difference between the two models. Let $L(\Omega_0)$ and $L(\Omega_1)$ be likelihoods from an unconstrained (i.e., a less restricted model) and a constrained (i.e., a nested model) each. Then the LR statistic is,

$$LR = -2\{log\ \ L(\Omega_1) - L(\Omega_0)\}, \tag{19}$$

and a significance of the LR is tested based on the Δdf (= $df_1 - df_0$). The LRT is also known as χ^2 difference test.

Since all statistical software reports the loglikelihood of a model, the LRT might be the most common test, and it same for testing partial factorial invariance. However, the LRT is also sensitive to sample size as χ^2 (Cheung & Lau, 2012) and it requires repeated model comparisons by estimating all models separately (Bollen, 1989; Cheung & Lau, 2012). Suppose testing partial metric invariance on six items under a single factor model. With the forward procedure, a nested model having an equally constrained factor loading across groups is compared to the baseline of the configural invariance model. The testing is performed on one-factor loading at a time; therefore, each partial metric invariance model is compared to the configural model and evaluated individually based on the LRT. Except for the reference variable, a total of five different model comparisons with the baseline model are necessary, and five LRT statistics are considered to identify partial metric invariance.

Wald Test

When it comes to testing partial invariance, the Wald test examines whether a parameter difference between groups is zero (e.g., $\lambda_{g1} - \lambda_{g2} = 0$) or not. The benefit of the Wald test is that, unlike the LRT, it does not require separate estimations of a baseline model and all partial invariance models. Multiple Wald tests are performed simultaneously by estimating a baseline model once with "MODEL CONSTRAINT" command in M*plus*.

The baseline model depends on the targeted parameter. For testing partial metric invariance which targets factor loadings, the baseline model is the configural invariance model. Except for a reference variable, the Wald test examines each equal constraint on the factor loading relative to the configural invariance, one at a time; however, all the Wald test results are reported simultaneously in the output of the baseline model. Likewise, the Wald test examines partial scalar invariance compared to the baseline model of metric invariance and tests partial strict invariance compared to the scalar invariance model.

While the Wald test has an analytical efficiency without a series of model estimations, it has a non-monotonically increased power issue. When the true population parameter is larger, the bigger Wald statistic for the estimate is expected; therefore, the power to detect the non-zero estimate is inclined to increase. However, previous studies found that the power of the Wald test was not monotonic compared to the tested population parameter size. The symptom was found in the regression model (Hauck Jr & Donner, 1977) and SEM (Chou & Bentler, 1990). In that case, detecting the larger non-invariant parameter might not always easier than the smaller non-invariant parameter in terms of measurement invariance testing. However, the non-monotonic issue has not been reported in the previous measurement invariance studies (Jung & Yoon, 2017; Kim et al., 2016).

On the other hand, previous measurement invariance studies used the Wald test for either the forward or backward procedure, and the Type I error rates (i.e., the false-positive rates) for using the Wald test were different depending on the direction of testing measurement invariance. Kim's et al. (2016) found the inflated Type I error rates of the Wald test for identifying partial invariance falsely with the backward procedure. In contrast, Jung and Yoon (2017) reported acceptable Type I error rates ($\alpha = 0.05$; Serlin, 2000) with the forward procedure.

Bias-Corrected Bootstrapping Confidence Interval (BCBS-CI)

Meade and Bauer (2007) showed the usefulness of confidence interval (CI) to test metric invariance by complementing other fit indices where the Type I errors were inflated. For example, if global fit indices indicate metric non-invariance, but the CI for the difference in a factor loading between groups is very narrow and almost close to zero, it implies that the actual difference is not large. In that case, the detected non-invariance might not be repeated with a different sample, especially with a smaller sample size. Like this, CIs will be valuable information for researchers to make a decision. Since the use of CI was proposed for testing partial measurement invariance, the follow-up CI studies have been continued. Recently, Mackinnon, Lockwood, and Williams (2004) suggested the bias-corrected bootstrapping CI (BCBS-CI), and its performance was evaluated in terms of testing measurement invariance (Cheung & Lau, 2012; Jung & Yoon, 2016).

Jung and Yoon (2016) examined the performance of the BCBS-CI for testing measurement invariance with several simulation conditions. The BCBS-CI had a lower Type I error of falsely detecting non-invariance than the MI method. However, when the sample size was small, its Type II error of misidentifying invariance for small non-invariance was higher than the MI. The pros and cons of BCBS-CIs might be balanced with perfect recovery rates. The perfect recovery rate indicates the degree to which tests detect true invariant parameters as invariant as well as identify true non-invariant parameters as non-invariant. In Jung and Yoon's (2016) study, the BCBS-CI was superior to other methods regarding the perfect recovery rates.

The BCBS-CI also has a computational advantage similar to the Wald test. Only a single estimation of a baseline model is required, without separate estimations for repeated partial invariance models. In addition to that, non-convergence errors are less likely to happen for the BCBS-CI, compare to other fit indices, because its baseline model is the least constrained model (Cheung & Lau, 2012).

However, the described test statistics for partial invariance are based on the normality assumption. However, it is not clear the studied characteristics of each test are similar or not when it comes to the zero-inflated data. Still, partial invariance and the related fit indices have not been studied regarding the zero-inflation, and the general strategies for the normal data are applied to the zero-inflated data. For example, Antoniadou, Kokkinos, and Markos (2016) used the traditional MGCFA and global fit indices to validate factorial invariance across grade and sex when high zero response rates were reported (min: 75%, Max: 92.7%).

This study tested the partial invariance of zero-inflated data through the two-part factor model and the MGCFA. Therefore, the performance of the general MGCFA regarding zero-inflation is also examined. In addition, the performance of the fit indices will be compared across the degrees of zero-inflation and the factor models (i.e., two-part factor model and MGCFA). In the next section, general statistical models to handle the zero-inflation are presented, and the two-part factor model is followed.

General Models to Handle Zero-inflation

Assume researchers want to know the effects of diverse variables on a severe disease. A survey asked the diagnosis of the disease and its stage or severity to know the current state. Most people will report that they have never been diagnosed with the disease, and their responses will be coded as zero. On the other hand, only a few people who have been diagnosed with it will report its severity, and their responses will be coded as Likert scale or continuous numbers. Thus, data will be inflated to zero because of the major responses (i.e., never diagnosed or experienced), and such a data distribution is called zero-inflation.

The zero-inflation separates the research question into two: 1) what are the effects of variables on the severity of the disease? and; 2) what makes people get a diagnosis or not. Health economy and medical expenditure studies have focused on these questions, and statistical models to handle the zero-inflation have been developed.

It is well known that a standard generalized linear model (i.e., linear regression analysis or ANOVA) is not proper for zero-inflated data (Duan et al., 1983). Alternative models are sample selection model, tobit model, multipart models (e.g., one-part, two-part, and four-part model), and zero-inflated (ZI) models (e.g., ZI Poisson and ZI negative binomial model). The alternative models are similar in some ways. The models treat zero and non-zero responses differently instead of considering all responses as a continuum by fitting different statistical models to zero and non-zero responses. In addition, different groups are computed depending on zero or non-zero responses. However, each alternative model has a different model assumption and the function to compute groups.

Although several models were developed, only a few models are currently used because of model weakness. For example, previous studies found that the sample selection model (i.e., adjusted tobit model) has an implausible assumption (i.e., normality assumption) and poor model estimation (Duan et al., 1984; Hay & Olsen, 1984). Therefore, other flexible models have been studied, and a two-part model and ZI models have been widely used for the applied study.

In this current study, the two-part model is mainly discussed because of its flexibility. Olsen and Schafer (2001) mentioned two-part model covers ZI Poisson model (Lambert, 1992). In addition, Kim and Muthén (2009) extended Olsen and Schafer's (2001) to SEM by suggesting two-part factor mixture modeling. Thus, the two-part model is not only more inclusive but also flexibly applied to SEM than ZI models.

Two-part Modeling

This section describes two-part modeling for generalized linear mixture model (Olsen & Schafer, 2001) and factor mixture model (Kim & Muthén, 2009).

Two-part Model

The definition and characteristics of the two-part model are referred to by Olsen and Schafer's (2001); however, they explained the model based on longitudinal data. This study focuses on two-part modeling with cross-sectional data; consequently, different notations are used.

The 'two-part' indicates a separation of zero-inflated responses into two data parts; binary-part and continuous-part. Let X is a p^{th} response of individual for $p = 1, \ldots, P$. Depending on whether X_p is zero or not, responses are redefined in each part (i.e., binary and continuous). The two-part is described with following:

$$U_p = \begin{cases} 1 \text{ if } X_p \neq 0 \\ 0 \text{ if } X_p = 0 \end{cases}$$

$$V_p = \begin{cases} f(X_p) & \text{if } X_p \neq 0 \\ \text{irrelevant} & \text{if } X_p = 0 \end{cases} \tag{20}$$

If X_p is non-zero (i.e., $X_p \neq 0$), U_p is coded as 1 in the binary-part, and V_p has a continuous value in the continuous-part where f is a monotonic increase function that will make V_p approximately normal. On the other hand, if X_p is zero (i.e., $X_p = 0$), U_p is coded as 0, and V_p is considered as a missing value.

Two parts have different characteristics; 1) The binary-part directly shows the degree of zero-inflation with zero or non-zero patterns by having a binomial model (e.g., logistic regression) (Xu et al., 2015) for U_p. Then effects of covariates on individual's probability of zero or non-zero response can be examined; 2) Continuous-part expresses the extent of an individual's experience by showing a continuous or count value which is conditional on the binary-part (i.e., if $U_p = 1$). Because V_p is usually skewed, log-linear models are typically used for the continuous-part to test relationships between covariates and V_p. Two parts' models are correlated each other.

The standard two-part model is a good approach to see the effects of covariates on the zero-inflated variable, but there are limitations. A two-part model analyzes the single zero-inflated variable, and the zero-inflated variable is treated as an outcome only. A multivariate two-part model has been developed to test multiple zero-inflated variables simultaneously (Brown, Ghosh, Su, & Taylor, 2015). However, the multiple zero-inflated variables still act as outcomes only, and the effect of the zero-inflated variable is not tested. By applying the two-part model to SEM, the role of the zero-inflated variable can be diverse rather than the outcome. For example, the zero-inflated variable can be either a predictor or an outcome variable, and the zero-inflated variable can take both roles simultaneously. Testing such a complex relationship regarding the zero-inflated variable is one of the benefits of SEM with two-part modeling.

Two-part Factor Model

Kim and Muthén (2009) demonstrated two-part modeling on SEM, especially on a factor mixture model with latent classes. The two-part factor mixture model, proposed by Kim and Muthén (2009), combines a two-part model and a factor mixture model. The current study defines two-part factor modeling, but we focus on multiple observed groups (e.g., demographic groups) rather than latent classes. Therefore, the original notation of the two-part factor mixture model is changed to a two-part factor model in this study.

Let X_g denote a $p \times 1$ vector of continuous responses for an individual in g^{th} subgroup, then a factor model is the same with Equation 4, as in MGCFA. For the two-part factor model, the factor model is decomposed into two parts, similar to the two-part model in Equation 20. Two-factor models having binary responses and continuous responses are respectively generated, and two factors are correlated. Factor models for each part are as follows:

$$X_g^* = \tau_{(U)g} + \Lambda_{(U)g}\xi_{(U)g} + \delta_{(U)g}$$
$$(U)_g = \begin{cases} 1 \; if \; X_g^* > \upsilon_g \\ 0 \; if \; X_g^* \leq \upsilon_g \end{cases}, \tag{21}$$

$$X_g = \tau_g + \Lambda_g\xi_g + \delta_g \tag{22}$$

Equation 21 represents a factor model for binary-part dealing with binary responses $(U)_g$. Let X_g^* denote a $p \times 1$ vector of latent continuous variates for g^{th} subgroup. Depending on threshold υ_g, observed binary responses $(U)_g$ have a $p \times 1$ vector showing zero or nonzero response. If X_g^* is greater than the threshold, $(U)_g$ is 1, otherwise $(U)_g$ is 0. X_g^* is a function of factor $\xi_{(U)g}$. In the factor model, $\tau_{(U)g}$ and $\delta_{(U)g}$ are a $p \times 1$ vector of intercepts and unique factors respectively, $\Lambda_{(U)g}$ indicates a $p \times m$ matrix for factor loadings, and $\xi_{(U)g}$ is a $m \times 1$ vector of factor scores.

Similar to Equation 20, X_g which is a $p \times 1$ vector of continuous responses in Equation 22 is also conditionally composes the factor model. As a monotonic increase function f, a logarithm function is usually used (Kim & Muthén, 2009) those for which $(U)_g$ is 1, otherwise X_g is treated as missing. These two factor models are correlated by having a factor correlation.

Because two-part modeling handles zero-inflated data, we expect the two-part factor model to outperform the traditional MGCFA to test factorial invariance on zero-inflated measures. The main goal of Kim and Muthén's (2009) Monte Carlo simulation study was to evaluate the multistage strategy to find a proper number of classes and factors for the two-part factor model. The simulation study mainly focused on the model evaluation; therefore, the correct detection of factorial invariance or non-invariance was not evaluated. Therefore, it was not enough to examine the model performances concerning factorial invariance, especially for observable subgroups. More in-depth studies are required to investigate the performance of the two-part factor model for testing factorial invariance. The current study is the early research discussing this issue by comparing the two-part factor model and the MGCFA when the observed variables are zero-inflated under various conditions.

Purpose of Studies

We aimed to compare the two-part factor model and the MGCFA about testing factorial invariance on continuous response. Study I used the empirical zero-inflated variables and focused on applying the two different models to testing the factorial invariance and evaluating the partial invariance model with test statistics. On the other hand, Study II was based on the simulated data. The two-part factor model and the MGCFA were fitted to the simulated zero-inflated data, and performances of the two models and the fit indices were compared based on how the factorial invariance was correctly tested. Overall research questions are as follows:

Q1: Do the two models perform differently in detecting violations of factorial invariance with zero-inflated data?

Q2: Will the performance of the two models be affected by the degree of zero-inflation?

Study I

The MGCFA was found as a popular model for handling the zero-inflation (Antoniadou et al., 2016) even after the two-part factor model has been introduced (Kim & Muthén, 2009). Therefore, the purpose of the first study was to apply the two-part factor model to testing factorial invariance for zero-inflated variables and to see its usefulness compared to the MGCFA's. In Study I, both the two-part factor model and the MGCFA are applied to empirical zero-inflated variables, and test procedures for the two models and factorial invariance results are compared to each other.

Study II

The second study aims to compare the performances of the two-part factor model and the MGCFA for metric and scalar invariance. We examined the performance of the two models about detecting the invariance and the non-invariance on the target parameter (i.e., factor loadings or intercepts). The different levels of factorial invariance were also manipulated. Test results of the two models were compared to the population parameters to see their absolute performance and compared to each other to see relative performances. Under each model, the Wald test and the BCBS-CIs were used to investigate partial invariance. The Type I error rate and power were outcome variables to evaluate the two models. The definitions of the outcome variables are described later in the method section.

CHAPTER III

STUDY I

Study I used empirical data. The two-part factor model and the MGCFA were applied to the same data, and their results regarding factorial invariance were compared. Across sex, factorial invariance in 10 continuous variables measuring bullying and victimization was tested. Participants, data, and analysis procedures for the two models are described next.

Method

Data

We used some of the delinquent variables for adolescents with zero-inflation from the Korean Youth Panel Survey (KYPS). The data was collected initially through the KYPS by the National Youth Policy Institute in 2003. A total of 3,449 students (male: 1,725) in the 2nd year of middle school responded to the survey.

The Study I focused on students' bullying and victimization experiences measured by 10 variables. Each 5-variable measured bullying or victimization by asking the frequencies of the different experiences over the past year. Three types of experiences were asked: physical, verbal, and general behaviors. Specifically, two variables asked physical behaviors ('*I have beaten up/been beaten up*' and '*I have robbed somebody/been robbed*'); two variables for asking verbal behaviors ('*I have mocked/been mocked*' and '*I have threatened/been threatened*'); and one variable measured the general bullying or victimization ('*I have bullied/been victimized somehow*'). Table 1 represents the descriptive statistics of the variables. The reported statistics considered all responses, including zeros without transformation; therefore, non-normality was noticeable by having high skewness and kurtosis. All variables had zero-inflation over 80%, and it impacted on the high non-normality.

Table 1

Descriptive Statistics of Variables across Sex

	Male (N = 1,725)					Female (N = 1,724)				
	Mean (SD)	min/Max	Skewness	Kurtosis	% of Zero	Mean (SD)	min/Max	Skewness	Kurtosis	% of Zero
					Bullying					
Mocking (B1)	1.78 (10.16)	0/100	8.71	78.94	80.40	0.66 (5.70)	0/100	14.34	229.52	91.00
Beating (B2)	0.51 (4.17)	0/100	18.49	401.57	87.70	0.19 (2.52)	0/100	36.15	1420.12	94.40
Threatening (B3)	0.29 (3.17)	0/100	23.26	643.99	94.60	0.12 (1.53)	0/50	25.49	755.99	97.00
Robbing (B4)	0.25 (3.12)	0/100	24.66	700.84	95.20	0.37 (3.43)	0/100	19.98	489.44	93.90
General (B5)	0.35 (4.12)	0/100	22.56	527.81	89.60	0.36 (1.55)	0/50	20.64	622.12	82.30
					Victimization					
Mocked (V1)	0.86 (6.45)	0/100	12.70	178.52	87.90	0.38 (4.21)	0/100	20.90	469.71	93.20
Beaten (V2)	0.20 (1.88)	0/45	17.97	367.64	94.30	0.07 (0.70)	0/19	17.97	398.55	97.60
Threatened (V3)	0.26 (1.92)	0/50	16.18	340.66	93.60	0.06 (0.70)	0/25	27.77	947.54	97.40
Robbed (V4)	0.38 (1.59)	0/25	8.48	95.17	87.00	0.16 (1.34)	0/50	30.62	1124.35	93.40
General (V5)	0.20 (3.58)	0/100	25.91	707.61	97.20	0.24 (3.29)	0/96	26.55	743.70	93.50

Note. B = Bullying variable; V = Victimization variable.

Analysis

Depending on the factor model, the zero-inflated 10 variables had different data formats for the analysis. In this section, different data formats and analysis procedures for the two-part factor model and the MGCFA were described. All analyses were conducted through M*plus*8.4 (Muthén & Muthén, 1998–2017) with a robust maximum likelihood estimator (MLR).

Data Preparation

For two-part modeling in M*plus*, DATA TWOPART command is necessary for handling the raw data. The command generates new binary and continuous variables representing the raw variables, and the characteristic of the original variables are separated into the binary and the continuous. If the raw value is zero, the new binary variable is coded as zero, and the new continuous variable is coded as missing. On the other hand, if the raw value is greater than zero, the new binary variable is coded as 1, and the new continuous variable has the log-transformed value of the original value. Therefore, each factor of the two-part separately constructs the factor model with new binary variables and the continuous variables in Equations 21 and 22. Without two-part modeling, either raw variables or manually transformed variables are used for the analysis depending on the degree of non-normality. Study 1 manually log-transformed the data with a constant (i.e., $f(X_p) = log(X_p + 1)$, $X_p > 0$) for fitting the MGCFA due to the severe zero-inflation of bullying/ victimization data.

Analysis Procedure

Measurement models of bullying and victimization were simultaneously tested across sex. Testing the configural invariance was conducted with the tentative reference variable by constraining the factor loading of the first variable as 1 because examining the global fit indices was the purpose at this step, and specific parameter estimates would not be considered. Since the metric invariance test, the focus of the study was extended to parameter estimates as well. Therefore, a reference variable of each construct was explored before testing the metric invariance model. It was to avoid the bias on the parameter estimate resulted from the non-invariant reference variable. Study I used the LRT to explore the possible invariant variables; particularly, Satorra-Bentler adjustment was applied to the LRT (hereinafter, S-B LRT) because the general LRT was not appropriate when the estimator was the MLR (Satorra & Bentler, 2010). A full metric invariance model was the baseline model constraining all factor ladings equally across sex. We examined the statistical difference between the baseline and the compared model, which freely estimated single factor loading. The comparison was conducted for a single factor loading at a time until all factor loadings were separately tested based on the same baseline model. Therefore, the S-B LRT tested the model difference on the freed factor loading given 1 Δdf the following backward procedure.

Identification

Parameter constraints are necessary on the latent variable to identify the baseline model without specifying the reference variable. The variance-covariance structure is identified by standardizing all factor variances (i.e., Φ_{ref} = I) for one of the groups (i.e., a reference group), and the factor variances of the other groups are freely estimated. In this case, all factor loadings are equally constrained across groups. There is no identification issue over the mean structure when it comes to the continuous variables. The factor means and the intercepts are not included in the analysis as 'no mean structure' is assumed (Muthén & Muthén, 1998–2017). However, the mean structure is necessary for the binary variables at the baseline model due to the thresholds of the binary variables. All thresholds are identically constrained across groups for identification of the mean structure regarding the binary variables, and the factor means of the reference group are fixed 0; however, the factor means for the other groups are freely estimated (Liu et al., 2016; Millsap & Yun-Tein, 2004). Study I followed this rule. First, all factor loadings were constrained for two groups (i.e., male and female). Second, the factor variances of males were fixed as 1, and they were freed in females. Third, factor means were fixed as 0 for both males and females if the factors were related to the continuous variables. These constraints corresponding to the continuous variables were the same for the two-part factor model and the MGCFA. However, the two-part factor model needed additional constraints due to the binary variables. All thresholds were equally constrained for the two groups, the factor means of the binary-part factors were fixed as 0 for male and freed for females.

Lopez Rivas, Stark, and Chernyshenko (2009) suggested a way to select a reference variable leading to improved power to detect the actual non-invariance. As following their way, we picked each factor's reference variable if the variable was invariant through the S-B LRT and had the highest factor loading at the baseline model. Based on the chosen reference variables, the metric and the scalar invariances were sequentially tested. When the full invariance was rejected, the partial invariance model was examined with the Wald test and the BCBS-CI. The two statistics tested the parameter differences between the two groups given the hypothesis assuming the parameter difference was zero (H0: $\lambda_{Male} - \lambda_{Female} = 0$ or $\tau_{Male} - \tau_{Female} = 0$). The detailed procedure for each model is described next.

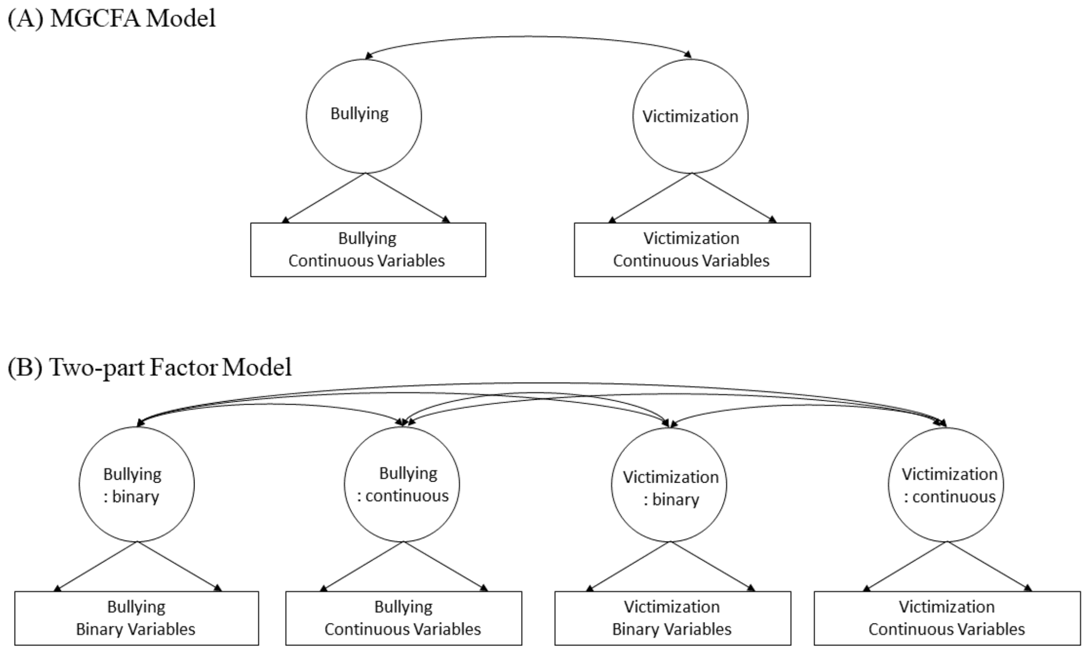

Figure 1 Analysis Factor Model for Bullying and Victimization.

MGCFA Analysis

The MGCFA represented a 2-factor model for bullying and victimization and examined the 10 continuous variables (Figure 1). Examining a reference variable through the S-B LRT was conducted first, and the metric and scalar invariance models were tested for the eight variables except for the reference variables. The Wald test with the Bonferroni correction (e.g., $\alpha = 0.05/8$ for the metric invariance model) and BCBS-CI of 99%. The MGCFA did not take the binary variables into account; therefore, the MGCFA was less demanding than the two-part factor model with fewer parameter estimates.

Two-part Factor Model Analysis

Bullying and victimization variables were separated into the binary-part factor and the continuous-part factor with the two-part modeling. Therefore, the factorial invariance in eight binary variables and eight continuous variables were tested except for four reference variables (Figure 1). For the two-part factor analysis for multiple groups, the following commands were required (Muthén & Muthén, 1998–2017):

1. TYPE = MIXTURE,
2. ALGORITHM = INTEGRATION.

The first command indicates mixture modeling to treat categorical latent variables, such as latent classes. Since this study aimed at observed two groups, we used 'known class' for the multi-group mixture modeling. Thus, results for each latent class were for each observed subgroup (Kim, Mun, & Smith, 2014; Muthén & Muthén, 1998–2017).

The second command aligns with TYPE = MIXTURE due to the demanding computation. Particular options under the algorithm command set details of the optimization to determine maximum likelihood estimates and specify the type of numerical integration and the number of integration points for computations. We used the default setting of M*plus*. In detail, a maximum likelihood estimator with robust standard error (MLR) was used with a numerical integration algorithm having 15 integration points for one dimension (Muthén & Muthén, 1998–2017).

Likewise, we tested the metric and the scalar invariance models sequentially after selecting the reference variables. The Wald test with the Bonferroni correction (e.g., $\alpha = 0.05/16$ for the metric invariance model) and BCBS-CI of 99%.

Outcome

Although the evaluation of the MGCFA model is possible based on the global fit indices with their criteria (RMSEA < 0.06, CFI ≥ 0.95, SRMR < 0.08; Hu & Bentler, 1999), there is no such fit index to assess the two-part factor model. Instead, information criteria (IC) indices, such as AIC (Akaike, 1987), BIC (Schwartz, 1978), or adjusted BIC (aBIC; Scolve, 1987), are considered. IC indices are based on the log-likelihood of the fitted model, and each IC index gives different penalties regarding a number of parameters or sample size (Nylund, Asparouhov, & Muthén, 2007). For example, AIC gives more penalties to the model having a larger number of parameters. On the other hand, BIC and aBIC give extra disadvantages to the larger sample size model in addition to the number of parameters. In general, one of the models having the lowest IC indices is selected among the competing models; that is, the lowest values are the most plausible given the penalties. However, the IC indices cannot evaluate the single model itself since there is no cutoff for them to select the model.

The direct model comparison between the two-part factor model and the MGCFA based on the global fit indices is not available because such fit indices are only reported for the MGCFA. In addition to that, the model comparison with the IC indices might be inappropriate because the two models differ in the number of parameters and the types of data being treated. The IC indices are known as inadequate for categorical data under SEM (Lai, 2020). The MGCFA treats all variables as continuous, but the two-part factor model manages both binary and continuous variables. Therefore, the global fit and the IC indices were used to evaluate different factorial invariance models (i.e., configural, metric, or scalar invariance model) in each factor analysis model, rather than comparing the two different factor analysis models. Also, the results from the two-part factor model and MGCFA were compared in three points: 1) the selection of the reference variables; 2) the supported invariance model, and; 3) invariant or non-invariant parameters in the model. In addition, the test results between the Wald test and the BCBS-CI were compared to see if there was any difference between the two statistics in identifying variables invariance.

Result

MGCFA

We initially tested the configural invariance model to examine whether or not male and female groups had identical factor structures. The global fit indices were acceptable (RMSEA = 0.03, CFI = 0.90, SRMR = 0.04, Table 5) by taking the combinational rule of Hu and Bentler's (1999) cutoff criteria (RMSEA < 0.06 , CFI ≥ 0.95, SRMR < 0.08). Before the metric invariance test, we selected the reference variables with the S-B LRT.

Table 2 represents the S-B LRT statistic of each variable. There were a total of 10 separate S-B LRTs to examine the equal factor loading constraint across sex. The second and fourth bullying variables (B2 and B4) had significant S-B LRT; that is, their factor loadings were statistically different between males and females. Among the variables representing the equal factor loadings across sex, we picked the first variables of bullying (B1) and victimization (V1) as the reference variables of each factor because their factor loadings were the highest at the baseline model (Kim et al., 2016; Lopez Rivas et al., 2009).

Table 3 shows the test statistics of factor loading and intercept invariance. Each Wald test and BCBS-CI examined the differences in the eight factor loadings between males and females at the configural invariance model. The Wald test and BCBS-CI indicated the non-invariance of B4 variable. Therefore, the partial metric invariance was considered instead of the full metric invariance. The partial metric invariance model had B4's factor loading estimated freely and constrained the other factor loadings equally across sex (RMSEA = 0.03, CFI = 0.91, SRMR = 0.05, Table 5). Next, the scalar invariance model on the seven intercepts was tested because the intercept of B4 was not considered due to its non-invariant factor loading. For all variables, equally constrained intercepts between male and female groups were rejected based on the Wald test and BCBS-CI; therefore, scalar invariance was not supported. In short, the partial metric invariance of the continuous variables was supported through the MGCFA; B4 had the non-invariant factor loading between the groups.

Two-part Factor Model

After testing the configural invariance model (AIC = 27831.12, BIC = 28476.43, aBIC = 28142.79, Table 6), the reference variables were searched. The two-part factor model tested each 10 binary and continuous variables; therefore, a total of 20 S-B LRTs were conducted. Table 2 presents the results on the reference variables. The third variables of bullying and victimization (binary B3 and V3) were chosen for the references for the binary variables. On the other hand, the third and fifth variables of bullying and victimization (continuous B3 and V5) were selected for the references of the continuous variables. The continuous B1 and V1 were chosen as the references in the MGCFA, however, they were considered to have non-invariant factor loadings with the two-part factor model.

Table 4 shows the test statistics for the factor loading and intercept invariance. The binary and continuous variables showed different results. In terms of the metric invariance, some of the factor loadings of the binary variables were non-invariant; however, all factor loadings of the continuous variables were invariant. Also, there were inconsistent test results between the Wald test and the BCBS-CI for the binary B1's factor loading. The Wald test indicated binary B1's factor loading was invariant for the male and female groups, but the BCBS-CI considered it non-invariant. The difference was because of the adjustment on the Type I error inflation. The Bonferroni correction ($\alpha = 0.05/16$) was only applicable to the Wald test; therefore, the BCBS-CI was likely to reject the null. To control the possible Type I error inflation, we accepted the Wald test result and equally constrained the factor loading of binary B1 for the two groups. However, the factor loadings of binary B4, B5, V5 were consistently non-invariant for two statistics, and they were estimated freely across sex. On the other hand, all continuous variables' factor loadings were considered invariant. In conclusion, the partial metric invariance in the binary variables and the full metric invariance in the continuous variables were supported. Next, we tested the scalar invariance of the continuous variables' intercepts only because intercepts of the binary variables were fixed to zero for the identification. All intercepts of continuous variables were considered invariant across sex, indicating the full scalar invariance of the continuous variables.

Taken together, the two different factor models led to different results for the continuous variables. First, different reference variables were chosen for each model. Second, the supported factorial invariance models were also varied. The MGCFA concluded partial metric invariance of the continuous variables. It also indicated the intercept non-invariance over all variables; that is, the observed mean differences between the groups might have been from the measurement bias rather than from the bullying and victimization constructs. Therefore, the group comparison with bullying and victimization variables was inappropriate based on the MGCFA. However, the two-part factor model concluded the full scalar invariance in the continuous variables and the partial metric invariance in the binary variables. With the two-part factor model, the group difference in bullying/victimization frequency was validated as they were derived from the differences in the bullying/victimization constructs.

Table 2

S-B LRT Results for Selecting Reference Variable

	B1	B2	B3	B4	B5	V1	V2	V3	V4	V5
MGCFA										
Continuous Variables	0.33+	**18.72**	2.22	**217.50**	9.09	0.05+	0.01	2.56	0.07	3.29
TPM										
Binary Variables	**11.49**	**10.83**	1.00+	**28.34**	**39.92**	0.58	1.53	2.62+	**7.85**	**40.25**
Continuous Variables	**7.80**	**2.92**	2.47+	0.18	0.19	**76.24**	3.18	0.39	0.02	0.15+

Note. + Reference variables having the highest factor loading; Bold indicates the significant and non-invariant factor loading given S-B LRT ($\chi^2_{(1)}$ = 3.84); B = Bullying variable; V = Victimization variable; TPM = Two-part factor model.

Table 3

Result of MGCFA on Factor loading and Intercept Invariance

	B1	B2	B3	B4	B5	V1	V2	V3	V4	V5
					Factor Loading					
Wald test	-	0.19	0.06	**-0.56**	-0.33	-	0.05	0.26	0.15	-0.28
BCBS-CI	-	[-0.21, 0.59]	[-0.36, 0.48]	**[-1.07, -0.04]**	[-0.75, 0.10]	-	[-0.68, 0.78]	[-0.25, 0.77]	[-0.76, 1.05]	[-0.78, 0.22]
					Intercept					
Wald test	-	**0.08**	**0.04**	**-**	**-0.07**	-	**0.04**	**0.06**	**0.08**	**-0.03**
BCBS-CI	-	**[0.05, 0.11]**	**[0.01, 0.06]**	-	**[-0.10, -0.04]**	-	**[0.02, 0.06]**	**[0.03, 0.08]**	**[0.05, 0.12]**	**[-0.06, -0.01]**

Note. B = Bullying variable; V = Victimization variable; Bold indicates significant test results that is non-invariant factor loadings given Wald test with Bonferroni correction (α = 0.05/8 for factor loading α = 0.05/7 for intercept) and BCBS-CI (99%)

Table 4

Result of Two-part Factor Model on Factor loading and Intercept Invariance

	B1	B2	B3	B4	B5	V1	V2	V3	V4	V5
					Factor Loading					
Binary Variables										
Wald test	0.20[A]	0.14	-	**-0.26**	**-0.44**	-0.07	0.16	-	0.11	**-0.86**
BCBS-CI	**[0.01, 0.39]**	[-0.02, 0.29]	-	**[-0.43, -0.08]**	**[-0.65, -0.23]**	[-0.31, 0.17]	[-0.20, 0.23]	-	[-0.05, 0.28]	**[-1.62, -0.11]**
Continuous Variables										
Wald test	-1.07	0.01	-	-0.18	-0.32	-0.47	0.0.63	0.19	0.13	-
BCBS-CI	[-3.05, 0.91]	[-1.05, 1.06]	-	[-1.30, 0.95]	[-1.00, 0.37]	[-1.95, 1.01]	[-0.39, 1.64]	[-0.80, 1.16]	[-0.74, 1.01]	-
					Intercept					
Continuous Variables										
Wald test	-0.05	-0.08	**-**	-0.41	-0.17	-0.17	-0.07	-0.20	-0.04	**-**
BCBS-CI	[-0.54, 0.45]	[-0.47, 0.31]	-	[-0.85, 0.04]	[-0.45, 0.12]	[-0.72, 0.38]	[-0.62, 0.47]	[-0.28, 0.67]	[-0.31, 0.39]	-

Note. B = Bullying variable; V = Victimization variable; Bold indicates significant test results that are non-invariant intercepts given Wald test with Bonferroni correction (α = 0.05/16 for factor loading and α = 0.05/8 for intercept) and BCBS-CI (99%); A indicates an inconsistent result between the Wald test and the BCBS-CI.

Table 5

Fit Indices of MGCFA

	Loglikelihood	No. of parameters	RMSEA	CFI	SRMR	AIC	BIC	aBIC
Configural Invariance	-9775.58	62	0.03	0.90	0.04	19675.16	19859.20	19859.20
Partial Metric Invariance : Factor loading of B4 was freed	-9840.07	55	0.03	0.91	0.05	19790.13	20128.15	19953.39

Note. B = Bullying variable; V = Victimization variable.

Table 6

Fit Indices of Two-Part Factor Model

	Log Likelihood	No. of parameters	AIC	BIC	aBIC
Configural invariance	-13810.56	105	27831.12	28476.43	28142.79
Metric Invariance Full Invariance of continuous variables Partial Invariance of binary variables : Factor loading of B4, B5, and V5 were freed	-13829.21	92	27842.42	28407.83	28115.51
Scalar Invariance Full Invariance of continuous variables	-13837.25	84	27842.51	28358.76	28091.85

Note. B = Bullying variable; V = Victimization variable.

CHAPTER IV

STUDY II

Study I demonstrated that how the two-part factor model and the MGCFA detected factorial invariance or non-invariance on the zero-inflated bullying/ victimization variables. Test results focused on the hypothesized factorial invariance on the variables. However, evaluating the two models to discuss which model more correctly examined the factorial invariance was not possible. Their performance can be compared to each other when the true factorial invariance or non-invariance pattern is known. Therefore, Study II conducted a Monte Carlo simulation study to test and compare the two models based on known population parameters. The data was generated based on the two-part factor model because we assumed that the two-part factor model might explain the highly zero-inflated variables better than the general factor model. Both the two-part factor model and the MGCFA were fitted, and their performance indicating the correct detection of factorial invariance or non-invariance were compared to each other. The Monte Carlo simulation study was conducted with M*plus*8.4 (Muthén & Muthén, 1998–2017).

Method

Simulation Conditions

Figure 2 represents the basis of generated data. The two-part factors (i.e., $F(U)$ and $F(X)$) had six variables (i.e., U_p or X_p) for each, and a factor correlation (i.e., $\phi_{F(u)F(X)}$) between them was 0.5. The factor correlation represented a linear relationship between the two-part processes. For example, a high correlation between the binary-part factor and the continuous-part factor indicates that individuals with a high propensity for an event might also have more frequencies for the event. From weak to strong, a variety of correlation coefficients (i.e., $\phi_{F(u)F(X)}$= [0.33, 0.90]) were reported in the empirical studies (Ferrer et al., 2016; McTernan & Blozis, 2014; Muthén, 2001). Study II set the moderate correlation as 0.5 based on the previous studies. To sum up, the generated data had 13 variables, and each part factor had 6-binary variables or 6-continuous variables.

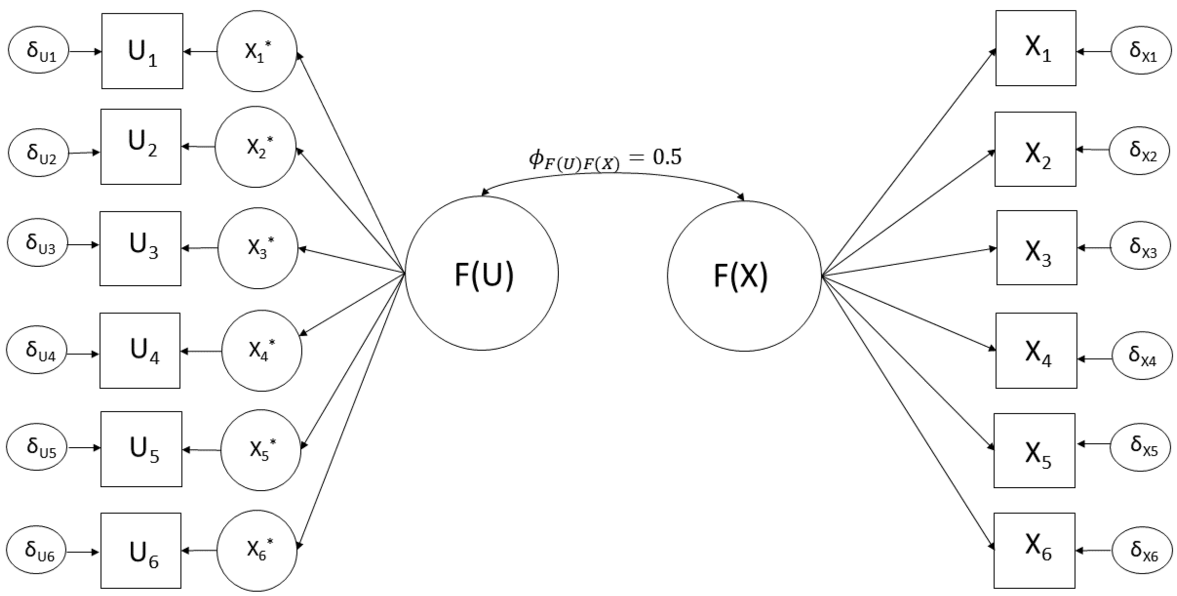

Figure 2 Generated two-part factor model.

Table 7 and Table 8 show population parameters. As the subgroups were fixed as two, some of the parameters were different across two group memberships (G1 and G2). Parameters from the binary-part factor model were identically generated across two groups (Table 1) because Study II examined the performance of the two models when it comes to the continuous variables. It was because MGCFA tests the continuous variables only. In contrast, population parameters for the continuous-part factor model were diversely manipulated for two groups (Table 2). The factor variance and the factor mean for G2 are slightly higher than G1 by taking the different factor score distributions across groups into account.

A total of 36 different conditions were generated varying three design factors: level of non-invariance (3) × sample size (3) × extent of zero-inflation (4). The number of replications was 1,000 for each condition. Each design factor was described in the next section. The simulation study was conducted separately for the metric invariance and scalar invariance models.

Level of Non-invariance

A single variable had non-invariant parameters among six continuous variables, and five variables were generated as invariant. For the metric invariance model, the non-invariant variable had a different factor loading (λ_2 in Table 2), but the corresponding intercept was set to be identical across groups. When it comes to the scalar invariance model, the intercept of the non-invariant variable (τ_2 in Table 2) had a different value across groups, but its factor loading was set to be identical.

The parameter difference between the two groups depended on the level of non-invariance: full invariance and two levels of partial invariance (i.e., small and large non-invariance). For the full invariance condition, both groups had the same factor loadings and intercepts. As the partial invariance, G2's parameters (λ_{x2} and τ_{x2}) were higher than G1's: 0.2 for the small non-invariance and 0.4 for the large non-invariance.

Table 7

Population Parameters for Binary-part Factor Model

	G1				G2			
	Percent of zero response							
	30%	50%	70%	90%	30%	50%	70%	90%
Threshold								
$\nu_{x_1^*}$	-.39	.13	.67	1.64	-.39	.13	.67	1.64
$\nu_{x_2^*}$	-.52	.0	.52	1.28	-.52	.0	.52	1.28
$\nu_{x_3^*}$	-.67	-.13	.39	1.04	-.67	-.13	.39	1.04
$\nu_{x_4^*}$	-.52	.0	.52	1.28	-.52	.0	.52	1.28
$\nu_{x_5^*}$	-.39	.13	.67	1.64	-.39	.13	.67	1.64
$\nu_{x_6^*}$	-.67	-.13	.39	1.04	-.67	-.13	.39	1.04
Factor loading								
$\lambda_{x_1^*} - \lambda_{x_6^*}$	.7				.7			
Unique variance								
$\varepsilon_{x_1^*} - \varepsilon_{x_6^*}$	.3				.3			
Factor variance-covariance								
φ_{11}	1.0				**1.3**			
φ_{12}	.5				.5			
Factor mean								
κ	.0				**.2**			

Note. Bold represents a higher value.

Table 8

Population Parameters for Continuos-part Factor Model

	G1	G2		
		Baseline	Small	Large
Factor loading				
λ_{x1}	.5	.5	.5	.5
λ_{x2}	.6	.6	**.8**	**1.0**
λ_{x3}	.7	.7	.7	.7
λ_{x4}	.6	.6	.6	.6
λ_{x5}	.5	.5	.5	.5
λ_{x6}	.7	.7	.7	.7
Intercept				
τ_{x1}	0.0	0.0	0.0	0.0
τ_{x2}	.2	.2	**.4**	**.6**
τ_{x3}	.3	.3	.3	.3
τ_{x4}	.4	.4	.4	.4
τ_{x5}	.5	.5	.5	.5
τ_{x6}	.6	.6	.6	.6
Unique factor variance				
ε_{x1}	.2			
ε_{x2}	.3			
ε_{x3}	.4			
ε_{x4}	.3			
ε_{x5}	.2			
ε_{x6}	.4			
Factor variance				
φ	1.0		**1.3**	
Factor mean				
κ	.0		**.2**	

Note. Bold represents a higher value.

Sample Size

The sample size is an important factor affecting convergence and fit of the factor analysis model (MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum, Widaman, Zhang, & Hong, 1999). The small ratio of the sample size over the number of observed variables might be a reason for non-convergence and unbiased parameter estimates. In addition, the effect of small sample size is more critical when the communality (i.e., a variance of common factor) is low. Therefore, Study II avoided the low and wide range of communalities to control the communality effect with the sample size and had communalities around 0.55 (MacCallum et al., 2001; MacCallum et al., 1999). The two groups had the same sample size. Small (N = 250), medium (N = 500), and large (N = 1,000) sample sizes were designed for each group, and consequently analyses were based on 500, 1,000, and 2,000 for the total sample sizes.

Extent of Zero-inflation

The current study referred to delinquency studies for reflecting realistic conditions of zero-inflated measures and for varying the level of zero-inflations. A high percentage of zero values (i.e., '*Never'* or '*No experience*') was found in bullying studies. Nansel (2001) studied the effects of various factors on bullying/ victimization; the average *'None*' rate was 55.7 %. On the other hand, You, Kim, and Kim (2014) reported overall high zero response rates, but the zero-inflations differed depending on the type of bullying. For example, 98.9% of 2nd year middle school students responded to '*No*' for sexual bullying, but 57.9% of them responded to '*No*' for cyber bullying. These examples have shown that the zero-inflation of delinquency measures is the common issue and overall high by showing some differences depending on the specific behaviors.

Study II reflected the empirical studies to set the simulation design. From low to high zero-inflation conditions, various zero-inflation levels were manipulated to see how the different extent of zero-inflation brings different performances of the two models. Therefore, four extents of zero-inflation (i.e., a percentage of zero responses) were designed: low (30%), moderate (50%), severe (70%), and extreme (90%). The observed variables had a slightly different percentage of zero-inflation depending on the threshold υ under the binary-part factor (Table 1). However, their average percentage was one of the four extents.

Data Analysis Procedure

The data adjustment processes were necessary to fit the MGCFA because the simulated data was based on the two-part factor modeling; the continuous variables were log-transformed, and zero responses were fixed as missing in the continuous-part model. The first process was to make the data without the transformation and the missing. Once the simulated data had been generated, the missing in the continuous-part was recoded as 0 to make the missing as the origin of the continuous values for the MGCFA. All continuous responses were transformed via the exponential function. For the second process, the dataset was log-transformed with a constant (i.e., $f(X_p) = log(X_p + 1), X_p > 0$) to lessen the data difference between the two models. Finally, the MGCFA used the data having 6-variables for a single factor model and one variable representing the group membership.

The two different factor models tested the metric and scalar invariance on the five variables except for the reference variable. For the metric invariance model, a baseline model was the configural invariance model with the freed factor loadings across groups. Each factor loading was equally constrained over groups (e.g., $\lambda_{x21} - \lambda_{x22} = 0$), therefore, a total of five pairs of factor loadings ($\lambda_{x2} - \lambda_{x6}$) were examined through the Wald test and BCBS-CIs. A Bonferroni correction was applied to the Wald test; hence, a strict *p*-value (p = 0.05/5 = 0.01) was considered for each test to control the inflated Type I error. Likewise, the BCBS-CIs were considered at 99% of the interval. For the scalar invariance model, a metric invariance model was a baseline model. All factor loadings were equally constrained, and intercepts were freely estimated. For the identification, the intercept of a reference variable was fixed to zero for two groups (i.e., $\tau_{x1} = 0$), and five pairs of the intercept were constrained (e.g., $\tau_{x21} - \tau_{x22} = 0$) across groups. The Wald test and the BCBS-CIs were applied in the same manner as the metric invariance test.

Outcome

We compared performances of the two-part factor model and the MGCFA in terms of a Type I error rate and power. In addition, we calculated the Type I error rates and the power with the Wald test's and the BCBS-CIs' results for the metric invariance model and the scalar invariance model. Therefore, the Type I error rates and the power for the two statistics (i.e., Wald test and BCBS-CIs) were also compared.

The Type I error rate was defined as a percentage of falsely detected non-invariance when the tested parameter was actually invariant. First, the Type I error was counted if any invariant parameter estimates were falsely identified as non-invariant for each iteration. Next, we calculated the Type I error rate by averaging Type I error over total replications at each condition. For generating the full invariance conditions, population parameters were identically manipulated for the two groups. Therefore, the Type I error rate was calculated for the 5-variable except a reference variable. On the other hand, the 4-variables were considered for the Type I error rates for partial invariance conditions because the condition had a reference variable and one non-invariant variable.

The power was defined as a correct detection of non-invariance when a population parameter was non-invariant. Because a single variable was manipulated to be non-invariant, the true positive non-invariant factor loading and the non-invariant intercept were counted for the power. We reported the power at the partial invariance conditions only because all population parameters were invariant across the groups at the full invariance conditions. The power for each condition was averaged over the total replications, similar to calculating the Type I error rate.

Result

Convergence Issue

For each condition, Monte Carlo simulation studies were iterated 1,000 times through the two-part factor model and the MGCFA. Overall, the two-part factor model was more related to the non-convergence because the zero-inflation resulted in the non-normality of the binary-part model and brought the missing of the continuous-part model. In addition, the two-part factor model was more computationally demanding than the MGCFA, and it was related to the non-convergence.

The non-normality and the small sample size were related to the problematic convergence. The two compared models had non-converged iterations for the conditions having extreme zero-inflation (90%). However, the two-part factor model showed a more severe convergence issue with the extreme zero-inflation when the sample size was small and medium (*N* = 250 and 500). The averaged non-converged iteration ratio from the two-part factor model was 0.70 with the small sample size and was 0.16 with the medium sample size for testing the metric invariance. The non-convergence was less severe for testing the scalar invariance; the averaged ratio of non-converged iterations was 0.43 with the small sample size and 0.04 with the medium sample size. The average non-convergence ratio for the other conditions was 0.004. The MGCFA also had the non-convergence when the conditions had the extreme zero-inflation and small or medium sample size. However, the MGCFA did not have severe non-convergence as much as the two-part factor model, and the non-convergence only happened when the metric invariance was tested. The averaged ratio of non-convergence iteration was 0.02 with the small sample size and 0.002 with the medium sample size for testing the metric invariance.

Type I Error Rates

We set the traditional criterion for the Type I error rate (i.e., $p < .05$) to evaluate the two models. Table 9 represents Type I error rates for the invariant variables through the two-part factor model and the MGCFA. Except for some conditions, both two models showed acceptable Type I error rates of less than 0.05.

In the metric invariance test, the two-part factor model and the MGCFA resulted in the opposite pattern of the Type I error rates across the zero-inflation extents. The two-part factor model showed the highest Type I error rates at the extreme extent (90%), and the MGCFA had the highest Type I error at the 30% of zero-inflation. More specifically, the averaged Type I error rate of the two-part factor model was 0.042. The Type I error rates were similar from 30% to 70% of zero-inflation conditions and acceptable levels (averaged 0.036), although they had minor differences at the third decimal place (e.g., 0.030 at 30%, $0.032_{\text{Wald}}/0.029_{\text{BCBS-CI}}$ at 50%, and $0.035_{\text{Wald}}/0.033_{\text{BCBS-CI}}$ at 70% when N =500 for small partial invariance). However, as Figure 3 shows, relatively higher Type I error rates were found at 90% of the zero-inflation, and it was more noticeable when the sample size was small. Under these conditions (i.e., N = 250, 90% of zero-inflation), the Type I error rates were higher than 0.05. Compared with the two-part factor model, the MGCFA had a smaller averaged Type I error (0.027), but the highest Type I error rates were the highest at the 30% of zero-inflation. In addition, the MGCFA generally had bigger differences between the smallest and the highest Type I error rates, compared with the two-part factor model having similar Type I error rates except for the small sample size conditions.

For testing the scalar invariance, the steep increase of the Type I error rates of the two-part factor model was not repeated. The averaged Type I error rate for the scalar invariance test was 0.041, similar to the metric invariance test. On the other hand, the MGCFA’s Type I error rates at the scalar invariance test were dissimilar compared with the pattern in metric invariance. The highest Type I error rates were still at 30% of zero-inflation when the sample size was small, likewise the metric invariance. However, as Figure 3 shows, the highest points of the Type I error rates are different for the moderate and large sample size; the Type I error rates are the highest at the moderate (50%) zero-inflation. The averaged Type I error rates from the MGCFA for test the scalar invariance test was 0.034, and it was also smaller than the two-part factor model.

The Wald test and BCBS-CIs did not show noticeable differences in each other when it comes to the Type I error rates. For example, for testing the metric invariance, the Type I error rate from the two-part factor model was 0.035 and 0.033 for the Wald test and the BCBS-CIs respectively when the sample size was medium (N = 500), the partial invariance was small, and the zero-inflation was 70%. The averaged Type I error rates with the two statistics were similar but different between the two factor models: $0.042_{\text{Wald}}/0.042_{\text{BCBS-CI}}$ through the two-part factor model, and $0.028_{\text{Wald}}/0.026_{\text{BCBS-CI}}$ with the MGCFA.

Although the binary-part factor model was generated with factorial invariance, and its result was not considered in Study II, we were able to look at Type I error rates for the metric invariance model of the binary-part. The pattern was the same as the continuous-part factor model. Except for some conditions with the small sample size and 90% of zero-inflation, the Type I error rates in the binary-part factor model were acceptable.

Power

In contrast to the Type I error rates, the power to correctly detect the non-invariant variable had obvious patterns across conditions; there were dramatic changes of the power depending on the extent of zero-inflation. Overall, the two-part factor model outperformed the MGCFA, and the two compared factor models showed the decreased powers when the zero-inflation was increased; the highest power was at 30% of the zero-inflation, and the lowest power was at 90% of the zero-inflation. Also, results from the Wald test and the BCBS-CIs were similar within the same factor model. Table 10 presents the results for the power.

More specifically, the power showed the four dominant patterns for design factors. Figure 4 represents the change of the power over design factors. We typically reported averaged results based on the metric invariance because patterns of the power were similar in both metric and scalar invariance tests. First, the two-part factor model showed a better power than the MGCFA. The average power from the two-part factor model was 0.495, and the average power through the MGCFA was 0.322. Second, the larger sample size positively affected the correct detection of non-invariance. The average power of the two-part factor model was 0.317 with the small sample, 0.504 with the medium sample, and 0.664 with the large sample size. The MGCFA showed a similar pattern but a smaller power than the two-part factor model; 0.175 with the small, 0.316 with the medium, and 0.476 with the large sample size. Third, the two different models detected the large partial invariance more easily than the small partial invariance. The average power of the two-part factor model was 0.335 to detect the small partial invariance and 0.655 to find the large partial invariance. The MGCFA had the same pattern; 0.183 for the small partial invariance and 0.461 for the large partial invariance. Lastly, the larger zero-inflation was more related to the false detection of the non-invariant parameters. Figure 4 shows how the power decreased as the zero-inflation increased from low (30%) to extreme (90%). Each averaged power at 30% and 90% was 0.788 and 0.108 for the two-part factor model and 0.657 and 0.010 for the MGCFA.

The powers between Wald test and BCBS-CIs were also similar to the Type I error rates. For example, when the sample size was medium (N = 500), the partial invariance was large, and the zero-inflation was 70%; the two-part factor model to detect metric non-invariance showed

0.665 and 0.660 power by using the Wald test and the BCBS-CIs respectively. For the same condition, the power using the MGCFA was 0.263 through the Wald test and 0.254 with the BCBS-CIs.

Table 9

Type I Error Rates for Invariant Variables

			Metric Invariance				Scalar Invariance			
			Wald Test		BCBS-CI		Wald Test		BCBS-CI	
Sample Size	Level of Invariance	% Of zero	2PM	CFA	2PM	CFA	2PM	CFA	2PM	CFA
250	Full	30	0.036	0.041	0.036	0.037	0.048	0.047	0.045	0.042
		50	0.042	0.017	0.038	0.017	0.043	0.036	0.037	0.031
		70	0.034	0.025	0.034	0.020	**0.058**	0.039	**0.053**	0.035
		90	**0.086**	0.018	**0.086**	0.018	**0.053**	0.033	**0.051**	0.029
	Small Partial	30	0.031	0.030	0.030	0.029	0.040	0.039	0.039	0.035
		50	0.033	0.015	0.032	0.015	0.035	0.032	0.030	0.028
		70	0.025	0.016	0.025	0.014	0.044	0.032	0.040	0.029
		90	**0.083**	0.017	**0.083**	0.017	**0.051**	0.029	0.048	0.028
	Large Partial	30	0.033	0.029	0.033	0.028	0.040	0.039	0.039	0.035
		50	0.036	0.014	0.034	0.013	0.035	0.033	0.030	0.028
		70	0.027	0.015	0.026	0.015	0.044	0.032	0.040	0.030
		90	**0.101**	0.018	**0.101**	0.017	0.046	0.029	0.044	0.024
500	Full	30	0.043	0.040	0.039	0.037	0.041	**0.052**	0.038	0.048
		50	0.040	0.037	0.035	0.034	0.038	**0.054**	0.036	**0.052**
		70	0.037	0.037	0.037	0.036	0.044	0.041	0.040	0.038
		90	**0.057**	0.010	**0.055**	0.010	0.044	0.016	0.042	0.015
	Small Partial	30	0.030	0.044	0.030	0.042	0.035	0.043	0.032	0.039
		50	0.032	0.035	0.029	0.032	0.029	0.044	0.029	0.042
		70	0.035	0.033	0.033	0.032	0.039	0.039	0.036	0.036
		90	0.048	0.009	0.047	0.008	0.041	0.015	0.039	0.015
	Large Partial	30	0.032	0.049	0.030	0.045	0.035	0.043	0.032	0.039
		50	0.033	0.036	0.030	0.032	0.029	0.044	0.029	0.042
		70	0.036	0.031	0.033	0.027	0.039	0.040	0.036	0.035
		90	0.050	0.009	0.048	0.006	0.041	0.015	0.039	0.015
1,000	Full	30	0.044	**0.051**	0.043	0.048	0.047	0.036	0.047	0.033
		50	0.042	0.037	0.041	0.033	**0.051**	0.043	0.050	0.040
		70	0.047	0.040	0.043	0.040	0.045	**0.051**	0.044	0.046
		90	0.045	0.013	0.043	0.013	0.045	0.026	0.043	0.019
	Small Partial	30	0.036	0.044	0.032	0.040	0.044	0.029	0.044	0.026
		50	0.038	0.031	0.035	0.027	0.046	0.038	0.045	0.035
		70	0.046	0.029	0.041	0.026	0.036	0.043	0.035	0.040
		90	0.042	0.011	0.040	0.011	0.038	0.023	0.036	0.018
	Large Partial	30	0.038	0.044	0.037	0.041	0.044	0.029	0.044	0.027
		50	0.038	0.029	0.036	0.027	0.046	0.038	0.045	0.035
		70	0.045	0.027	0.041	0.024	0.036	0.042	0.035	0.040
		90	0.043	0.011	0.041	0.011	0.038	0.022	0.036	0.018

Note. 2PM = Two-part factor model; CFA = Multi-group confirmatory factor model. Type I error rates above 0.05 are in bold.

Table 10

Power for a Non-invariant Variable

Sample Size	Level of Invariance	% Of zero	Metric Invariance: Wald Test 2PM	Metric Invariance: Wald Test CFA	Metric Invariance: BCBS-CI 2PM	Metric Invariance: BCBS-CI CFA	Scalar Invariance: Wald Test 2PM	Scalar Invariance: Wald Test CFA	Scalar Invariance: BCBS-CI 2PM	Scalar Invariance: BCBS-CI CFA
250	Small Partial	30	0.304	0.183	0.300	0.176	0.437	0.156	0.430	0.151
		50	0.191	0.081	0.185	0.079	0.293	0.069	0.284	0.067
		70	0.087	0.028	0.084	0.026	0.145	0.036	0.141	0.033
		90	0.017	0.001	0.017	0.001	0.034	0.007	0.032	0.005
	Large Partial	30	**0.872**	0.650	**0.870**	0.646	**0.980**	0.726	**0.979**	0.722
		50	0.646	0.358	0.640	0.352	**0.904**	0.340	**0.902**	0.326
		70	0.328	0.098	0.324	0.097	0.622	0.110	0.610	0.108
		90	0.105	0.009	0.098	0.009	0.103	0.020	0.097	0.020
500	Small Partial	30	0.636	0.394	0.629	0.389	0.783	0.358	0.773	0.350
		50	0.401	0.182	0.399	0.176	0.549	0.156	0.544	0.150
		70	0.210	0.053	0.200	0.051	0.283	0.052	0.274	0.047
		90	0.055	0.005	0.053	0.003	0.060	0.013	0.059	0.011
	Large Partial	30	**0.994**	**0.950**	**0.994**	**0.948**	**1.000**	**0.968**	**1.000**	**0.963**
		50	**0.947**	0.687	**0.946**	0.681	**0.999**	0.659	**0.999**	0.652
		70	0.665	0.263	0.660	0.254	**0.914**	0.244	**0.911**	0.242
		90	0.141	0.009	0.140	0.009	0.217	0.037	0.209	0.033
1,000	Small Partial	30	**0.927**	0.778	**0.924**	0.767	**0.986**	0.754	**0.985**	0.738
		50	0.732	0.393	0.725	0.381	**0.927**	0.320	**0.923**	0.313
		70	0.401	0.119	0.396	0.114	0.642	0.104	0.633	0.100
		90	0.082	0.007	0.081	0.007	0.113	0.025	0.110	0.023
	Large Partial	30	**1.000**	**0.999**	**1.000**	**0.999**	**1.000**	**1.000**	**1.000**	**1.000**
		50	**1.000**	**0.959**	**1.000**	**0.959**	**1.000**	**0.962**	**1.000**	**0.958**
		70	**0.926**	0.548	**0.923**	0.535	**0.998**	0.516	**0.998**	0.505
		90	0.252	0.029	0.249	0.026	0.499	0.071	0.492	0.065

Note. 2PM = Two-part factor model; CFA = Multi-group confirmatory factor model; Acceptable power ($p \geq 0.80$) is in bold for each condition.

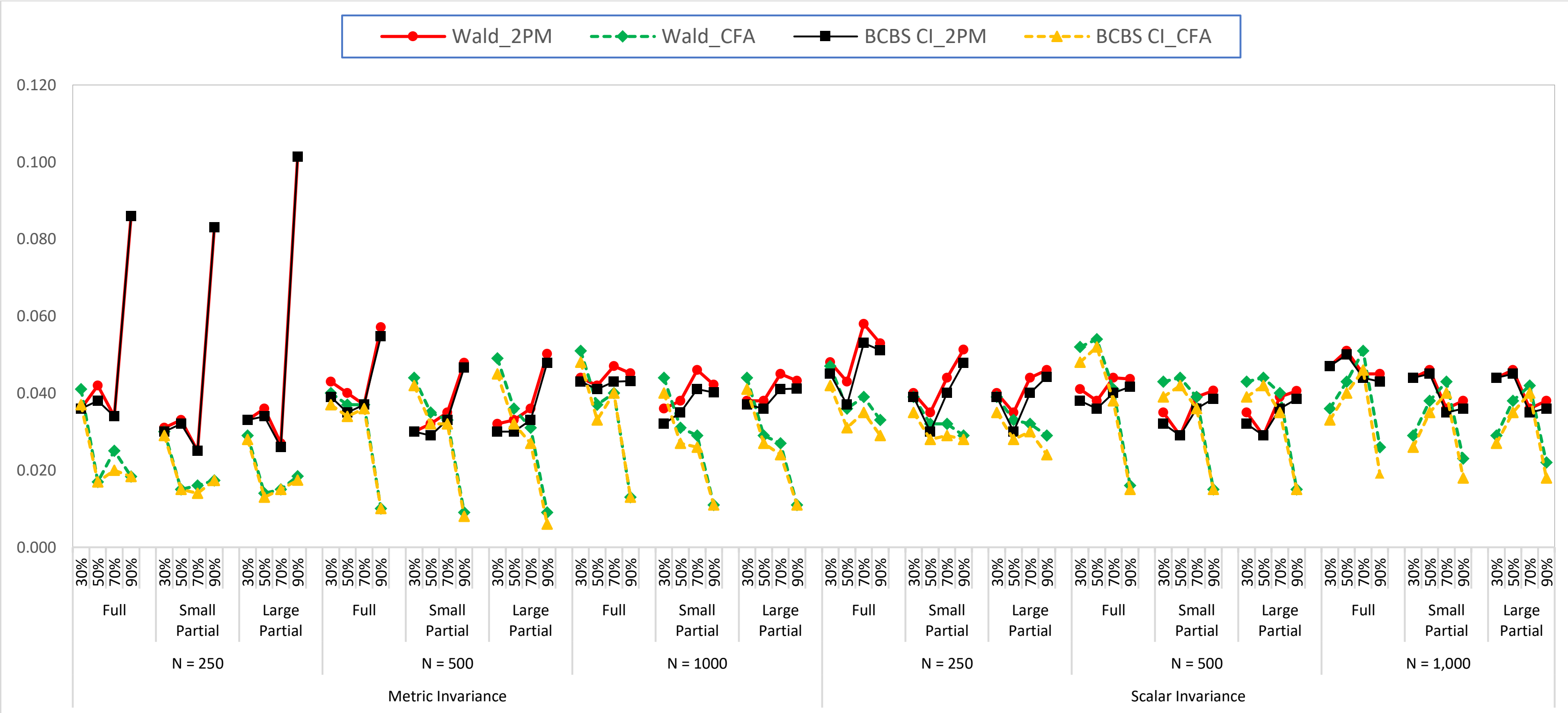

***Figure** 3* Type I Error Rates across Conditions.

Note. 2PM = Two-part factor model; CFA = General confirmatory factor model; Percentage (%) indicates the zero-inflation percent of the data; Full = Full invariance; Small = Small partial invariance; Large = Large partial invariance.

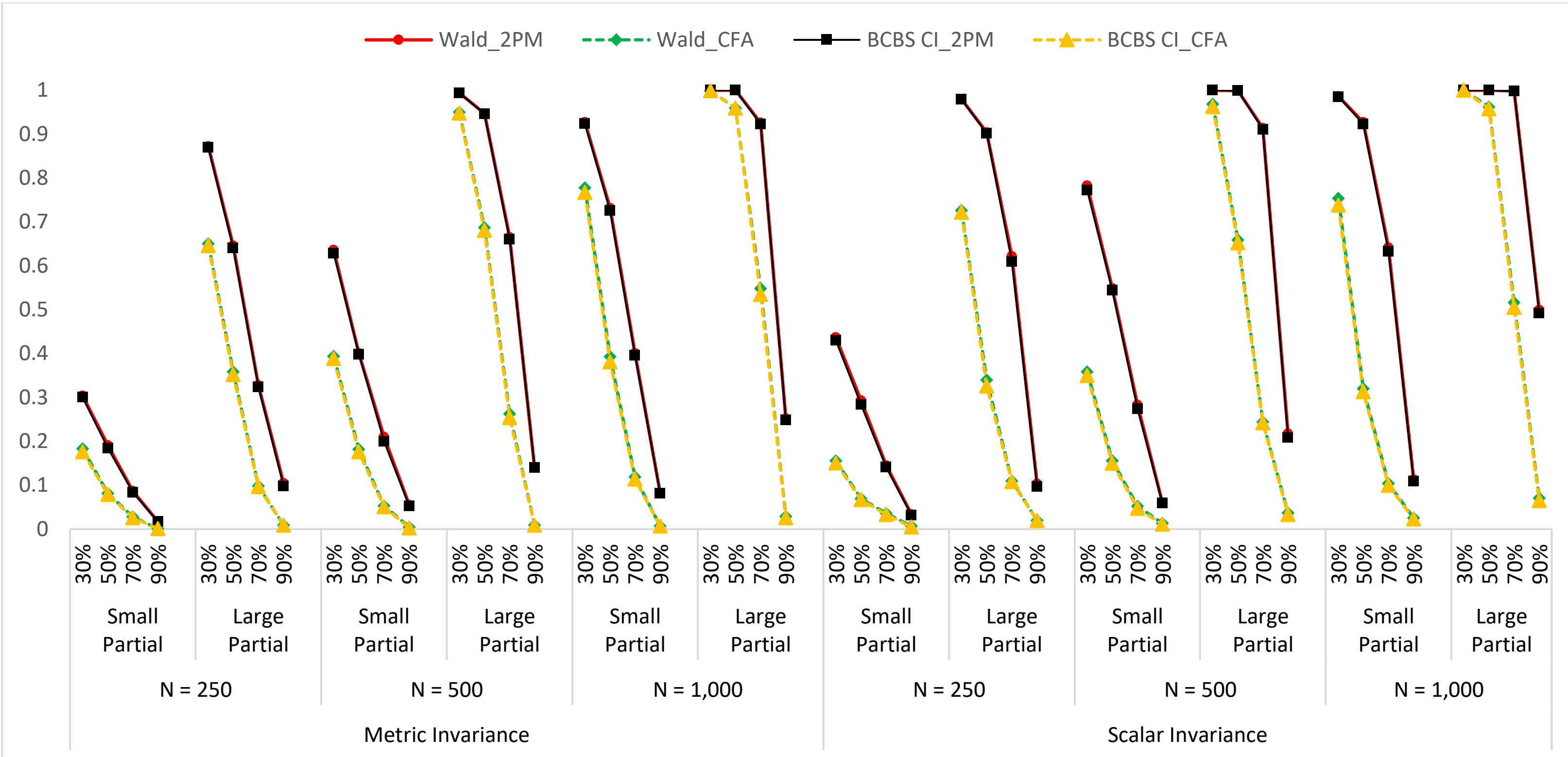

Figure 4 Power across Conditions.
Note. 2PM = Two-part factor model; CFA = General confirmatory factor model; Percentage (%) indicates the zero-inflation percent of the data; Small = Small partial invariance; Large = Large partial invariance.

Effect of Design Factors

We tested the effects of the design factors on the Type I error rate and the power with the mixed ANOVA model: (a) between-subject factors: sample size (2) × level of non-invariance (3) × extent of zero-inflation (4); (b) a within-subject factor: factor model (2).

Table 11 represents the result of the Type I error rates. Looking at the metric invariance test results, the main effects of between factors were not significant. However, the between factors interacted with the factor models and resulted in significant effects on the Type I error rates. For example, the significant interaction effect between the sample size and the factor model showed that the sample size effects on the Type I error rates were different in terms of the two factor models. The biggest effect size was from the interaction effect between the extent of zero-inflation and the factor model (η_p^2 = 0.865). On the other hand, the interaction between the level of non-invariance and the factor model was not significant. It indicated that the different levels of non-invariance had subtle effects on the Type I error rates when the zero-inflated variables were analyzed with either the two-part factor model or the MGCFA. The patterns of design factor effects were a bit different in the scalar invariance test. The between effects of level of non-invariance and extent of zero-inflation were significant, and the interaction effect between the sample size and the factor model was not significant.

When it comes to power, we identified the four dominant patterns from the design factors. The ANOVA result in Table 12 also shows the significant main effects of the between factors and a within factor on the power. The extent of zero-inflation had the biggest effect size (η_p^2 = 0.836 for metric invariance, η_p^2 = 0.883 for scalar invariance).

Table 11

ANOVA Results for Type I Error Rates

Source	Metric Invariance				Scalar Invariance			
	df	F	*p*	η_p^2	df	F	*p*	η_p^2
Between-subject Effects								
N	2	0.089	0.915	0.006	2	0.559	0.578	0.038
LN	2	1.239	0.305	0.081	2	**10.014**	**0.001**	**0.417**
EZ	3	1.446	0.251	0.134	3	**6.078**	**0.003**	**0.394**
Error	28				28			
Within-subject Effects								
FM	1	**107.570**	**< .01**	**0.793**	1	**15.569**	**< .01**	**0.357**
FM x N	2	**13.699**	**< .01**	**0.495**	2	3.026	0.065	0.178
FM x LN	2	0.375	0.691	0.026	2	0.224	0.801	0.016
FM x EZ	3	**60.036**	**< .01**	**0.865**	3	**10.572**	**< .01**	**0.531**
Error	28				28			

Note. η_p^2 = Partial eta square; N = sample size; LN = Level of Non-invariance; EZ = Extent of Zero-inflation; FM = Factor Model; Bold indicates statistically significant results.

Table 12

ANOVA Results for Power

Source	Metric Invariance				Scalar Invariance			
	df	F	*p*	η_p^2	df	F	*p*	η_p^2
Between-subject Effects								
N	2	**12.140**	**< .01**	**0.588**	2	**16.701**	**< .01**	**0.663**
LN	1	**31.018**	**< .01**	**0.646**	1	**52.744**	**< .01**	**0.756**
EZ	3	**28.836**	**< .01**	**0.836**	3	**42.906**	**< .01**	**0.883**
Error	17				17			
Within-subject Effects								
FM	1	**59.554**	**< .01**	**0.778**	1	**48.469**	**< .01**	**0.740**
FM x N	2	0.466	0.635	0.052	2	0.119	0.888	0.014
FM x LN	1	0.874	0.363	0.049	1	0.161	0.693	0.009
FM x EZ	3	2.491	0.095	0.305	3	2.620	0.084	0.316
Error	17				17			

Note. η_p^2 = Partial eta square; N = sample size; LN = Level of Non-invariance; EZ = Extent of Zero-inflation; FM = Factor Model; Bold indicates statistically significant results.

CHAPTER V

CONCLUSION AND DISCUSSION

Discussion

The current study started with the concern that only a few studies discussed the factorial invariance of zero-inflated variables (Antoniadou et al., 2016; Argyriou, Um, Wu, & Cyders, 2020), in contrast to many studies that have discussed naturally zero-inflated behaviors and validated the related measures. The lack of studies demonstrating the two-part factor model for the zero-inflated variables might be the reason for this matter. Based on the result of this study, applying the MGCFA to the zero-inflated variables might result in the bias of the related studies. In this respect, we suggested a two-part factor model for the zero-inflated data to test the factor invariance.

The MGCFA is a typical approach to test factorial invariance under the SEM (Meredith, 1993; Millsap, 2012). Previous studies have reflected the various methodological perspectives on the MGCFA, and the non-normality issue is also covered with the MGCFA while the helpful methods are applied together. For example, the non-normality is taken into account with the robust estimator (Brace & Savalei, 2017; Chen, Wu, Garnier-Villarreal, Kite, & Jia, 2020; Liu et al., 2016), or with the mixture modeling (Kim & Muthén, 2009) when the MGCFA is the analysis model. However, as Muthén (1989) discussed, a lot of zero observations might mislead the conclusion when the MGCFA is applied to the data because it treats all zero values as the origin. The data transformation of the zero-inflated variables can be aligned with the MGCFA to lessen the non-normality; however, the inflation at a certain value still exists. Using the two-part modeling to separate many zeros from the other continuous values might be more helpful. In that sense, Study I and Study II showed the usefulness of two-part factor model for testing factorial invariance of the zero-inflated data compared to the MGCFA.

Study I showed the use of the two-part factor model for testing the factorial invariance on the zero-inflated variables. We examined the factorial invariance on bullying and victimization variables having zero-inflation across sex. The MGCFA treated the zeros as the origin of continuous values; hence, all values were analyzed as the continuous variables under the corresponding single-factor structure. Therefore, bullying and victimization variables constructed each of the bullying and victimization factors. On the other hand, the two-part factor model separated zeros and the other values into two parts; all zeros were analyzed through the binary-part factor, and the other continuous values were managed with the continuous-part factor. Therefore, bullying and victimization were represented with each two-part factor model. The results of the MGCFA and the two-part factor model were compared in terms of the results of the S-B LRT, the Wald test, and the BCBS-CIs which frequently used test statistics in testing the factorial invariance (Cheung & Lau, 2012; Kim et al., 2016; Meade & Bauer, 2007; Jung & Yoon, 2016). We only compared the test results of the continuous variables, and they were dissimilar across models. The chosen reference variables were different, and the supported factorial invariance models were also varied. The MGCFA supported the partial metric invariance for the bullying and victimization factor models. Therefore, the sex comparison was not recommended based on the MGCFA. The two-part factor model also detected the non-invariance on the variables across sex; however, the non-invariance was on the binary variables measuring the binary-part factors, and the continuous variables under the continuous-part factors had the full scalar invariance. In other words, when the binary-part controlled the zero values, the sex comparison in terms of the continuous values was appropriate. There could be several reasons for the different results of the two models. In general, the traditional model such as ANOVA is known as inadequate for the zero-inflated data; thererfore, researchers recommend the ZI models or the two-part model to analyze such data (Duan et al., 1983). In this context, we hypothesized that the results of the two models differed because the typical factor model might not have been adequate for handling zero-inflation and conducted Study II to investigate it.

While Study I empirically showed the different test results of the two models, Study II examined which data characteristics affected the different results between the two models and how much the characteristics influenced the test results with the simulation study. The level of non-invariance, the sample size, and the extent of zero-inflation were manipulated to examine their effects on testing the metric and scalar invariance and to compare performances between the two different factor models. Generally, the ANOVA results showed the importance of model selection to test factorial invariance on the zero-inflated variables. Applying different factor models resulted in statistically varied Type I error rates and power. It indicates that the choice of the factor model might lead to different test results. It is in line with previous studies discussing specialized models (i.e., ZI models or two-part model) is necessary to handle the zero-inflation rather than the general models (Deb, Hall, Trivedi, & Hall, 2002; Duan et al., 1983; Lambert, 1992; Xu et al., 2015).

Except for some of the conditions, both the MGCFA and the two-part factor model showed the acceptable Type I error rates (i.e., $\alpha < 0.05$) in Study II. The MGCFA generally showed smaller Type I error rates than the two-part factor model, especially it was noticeable when the sample size was small, and the zero-inflation was 90%. However, it is difficult to conclude that the MGCFA is a better model for analyzing zero-inflated data in those conditions because the MGCFA showed lower power than the two-part factor model in any conditions. Based on the Study II result, when the extent of zero-inflation increased, the MGCFA had difficulty identifying the accurate partial invariance model. The correct model detection was harder when the sample size was decreased. However, its power was appropriate to test large partial invariance when the data had 30% of zero-inflation, and the sample size was the medium or large sample. The MGCFA's power was also acceptable when the zero-inflation was 50%, but large sample size was required. The possible source of the insufficient power of MGCFA might be from the approach to the zeros. The MGCFA considers the zero as the origin of the continuous values and keeps the original zero-inflation of the data, in contrast to the two-part factor model lessening the zero-inflation by separating the zeros into the binary-part factor model.

In conclusion, the current study showed that the MGCFA did not perform well to the zero-inflation as its power became lower with the increased zero-inflation. However, it is difficult to say the results were inconsistent with the previous studies that found the robust performance of the MGCFA to the non-normality. They showed adequate Type I error rates and power of the MGCFA across various non-normal conditions (Brace & Savalei, 2017; Chen et al., 2020). It is because our study and the prior research had differences in the approaches for testing the factorial invariance model and the type of non-normal data. The general definitions of the Type I error rate and the power were similar. The Type I error was to identify the factorial non-invariance falsely, and the power was to detect the true factorial non-invariance. However, the target models were different. We investigated the factorial invariance model by testing individual parameters based on the difference of 1 degree of freedom ($\Delta df = 1$). On the other hand, the previous studies focused on testing the multiple parameters together. The used statistics were also varied. Unlike our study that used the Wald test and the BCBS-CI for testing factorial invariance, previous studies used the LRT. More importantly, the zero-inflation might need to be considered separately from the typical non-normality. The zero-inflation, a special case of non-normality, shows that a lot of responses gather at the smallest value. Thus, it is a bit different pattern from the widely studied non-normal data, which has gradual decreases on both sides of the peak of the distribution. These differences make it hard to compare the present study to the previous studies directly. Future studies for testing partial factorial invariance over the zero-inflation are required to compare more results and find a better strategy for this type of non-normality.

Lastly, we found that the two-part factor model resulted in inadequate Type I error rates when the sample size was small, and the zero-inflation was extreme. With a sample size of 250 and zero-inflation of 90%, the Type I error rates for identifying full, small, and large partial metric invariance models were 0.087, 0.083, and 0.101, respectively. Because of the high rates of non-convergence at these particular conditions, we considered the non-convergence under these conditions could represent the possible reason for the unacceptable performance of the two-part factor model. Nevitt and Hancock (2004) discussed the association of non-convergence, improper solutions, and the poor Type I error for testing SEM. The non-convergence was likely to occur from the small sample size, and the Type I error rates within the corresponding simulation condition were also high. Although there was no direct relationship between the non-convergence and the Type I error rates, models with too small sample size were difficult to converge, and if they could be converged, the probability of Type I error would be greater than with larger sample sizes. In addition to the small sample size (Chen et al., 2020; Liu et al., 2016; Rhemtulla, 2012), high missingness and non-normality (Chen et al., 2020) were prone to the non-convergence. Because the zero values were recoded as missing in the continuous-part model, we contemplated that the missingness was also related to the higher Type I error issue within the conditions resulting in the high non-convergence. In the meaning that the model complexity is related to the non-convergence, the mixture modeling showed more non-convergence than the typical model (Henson, Reise, & Kim, 2007). That is, the two-part factor model for multiple groups, which is a combination of factor mixture modeling and two-part modeling, would have more difficulties to converge than the MGCFA, and it might be related to higher Type I error. In sum, the high Type I error rates of the two-part factor model were related to these reasons: mixture modeling with small sample size and high missingness due to extreme zero-inflaion. Therefore, we recommend the two-part factor model with a large enough sample size to avert non-convergence and to control Type I error for testing factorial invariance of the zero-inflated data.

Limitation and Future Study

We focused on zero-inflation as one type of non-normality and mainly examined its effects on testing partial factorial invariance. The limited design factors were considered in the current study; however, future studies can extend the current study for further investigation.

First, we only manipulated different parameters regarding the continuous-part factor model but assumed the factorial invariance of the binary-part factor model across groups. Therefore, the future study can differentiate the binary-part factor model across groups and compare the two models to see any confounding effect between the binary-part and continuous-part factors.

Second, various factor correlations can be manipulated between the binary-part factor and the continuous-part factor. We controlled the factor correlation as the moderate size. However, different factor correlation levels can clarify the effect of a correlational relationship between the two-part factors on the performance of the models.

Lastly, the missingness might be another characteristic of the zero-inflated variables in the future study. The two-part factor model treats zeros values as the zeros in the binary-part factor model and as missing in the continuous-part factor. However, the missingness can be added for both factor models of the two-part if there are true missing responses. For example, the survey asking the organizational turnover is sensitive to answer. The employees might not want to answer the questions rather than responding to them as they are not likely to change their jobs (i.e., zero answers). The different levels of missingness and the zero-inflation depending on the field might be helpful to get the practical strategy regarding the zero-inflated data for specific settings.

REFERENCES

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

Antoniadou, N., Kokkinos, C. M., & Markos, A. (2016). Development, construct validation and measurement invariance of the Greek cyber-bullying/victimization experiences questionnaire (CBVEQ-G). *Computers in Human Behavior, 65*, 380-390. doi:10.1016/j.chb.2016.08.032

Argyriou, E., Um, M., Wu, W., & Cyders, M. A. (2020). Measurement invariance of the UPPS-P impulsive behavior scale across age and sex across the adult life span. *Assessment, 27*(3), 432-453.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin, 107*(2), 238-246.

Bentler, P. M. (1995). *EQS structural equations program manual*: Multivariate software.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: John Wiley & Sons.

Brace, J. C., & Savalei, V. (2017). Type I error rates and power of several versions of scaled chi-square difference tests in investigations of measurement invariance. *Psychol Methods, 22*(3), 467-485. doi:10.1037/met0000097

Brown, E. C., Catalano, R. F., Fleming, C. B., Haggerty, K. P., & Abbott, R. D. (2005). Adolescent substance use outcomes in the Raising Healthy Children project: a two-part latent growth curve analysis. *J Consult Clin Psychol, 73*(4), 699-710. doi:10.1037/0022-006X.73.4.699

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

Chen, P. Y., Wu, W., Garnier-Villarreal, M., Kite, B. A., & Jia, F. (2020). Testing Measurement Invariance with Ordinal Missing Data: A Comparison of Estimators and Missing Data Techniques. *Multivariate Behav Res, 55*(1), 87-101. doi:10.1080/00273171.2019.1608799

Cheung, G. W., & Lau, R. S. (2012). A Direct Comparison Approach for Testing Measurement Invariance. *Organizational Research Methods, 15*(2), 167-198. doi:10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-fit Indexes for Testing Measurement Invariance. *Structural equation modeling, 9*(2), 233-255.

Chou, C.-P., & Bentler, P. M. (1990). Model modification in covariance structure modeling: A comparison among likelihood ratio, Lagrange multiplier, and Wald tests. *Multivariate Behavioral Research, 25*(1), 115-136.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (J. Cohen & P. Cohen Eds. 2nd ed.). Hillsdale, N.J.: L. Erlbaum Associates.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29.

Deb, P., Hall, C., Trivedi, P. K., & Hall, W. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of health economics, 21*(4), 601-625.

Deb, P., & Holmes, A. M. (2000). Estimates of use and costs of behavioral health care: a comparison of standard and finite mixture models. *Health Economics, 9*, 475-789.

Deb, P., & Trivedi, P. K. (2002). The structure of demand for health care: Latent class versus two-part models. *Journal of health economics, 21*(4), 601-625.

Duan, N., Manning, J. W. G., Morris, C. N., & Newhouse, J. P. (1983). A Comparison of Alternative Models for the Demand for Medical Care. *Journal of Business and Economic Statistics, 1*, 115-126.

Duan, N., Willard G. Manning, J., Morris, C. N., & Newhouse, J. P. (1984). Choosing between the Sample-Selection Model and the Multi-Part Model. *Journal of Business & Economic Statistics, 2*(3), 283-289.

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

Ferrer, E., Conger, R. D., & Robins, R. W. (2016). Longitudinal Dynamics of Substance Use and Psychiatric Symptoms in Count Data with Zero Inflation. *Multivariate Behav Res, 51*(2-3), 279-295. doi:10.1080/00273171.2016.1144501

Geiser, C. (2013). *Data analysis with Mplus*. New York: The Guilford Press.

Hay, J. W., & Olsen, R. J. (1984). Let Them Eat Cake: A Note on Comparing Alternative Models of the Demand for Medical Care. *Journal of Business & Economic Statistics, 2*(3), 279-282.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting Mixtures From Structural Model Differences Using Latent Variable Mixture Modeling: A Comparison of Relative Model Fit Statistics. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(2), 202-226. doi:10.1080/10705510709336744

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144.

Howell, D. C. (2012). *Statistical Methods for Psychology* (8 ed.): Cengage Learning.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling, 6*(1), 1-55. doi:10.1080/10705519909540118

Jung, E., & Yoon, M. (2016). Comparisons of Three Empirical Methods for Partial Factorial Invariance: Forward, Backward, and Factor-Ratio Tests. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 567-584. doi:10.1080/10705511.2015.1138092

Jung, E., & Yoon, M. (2017). Two-Step Approach to Partial Factorial Invariance: Selecting a Reference Variable and Identifying the Source of Noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(1), 65-79. doi:10.1080/10705511.2016.1251845

Kaysen, D., Atkins, D. C., Simpson, T. L., Stappenbeck, C. A., Blayney, J. A., Lee, C. M., & Larimer, M. E. (2014). Proximal relationships between PTSD symptoms and drinking among female college students: results from a daily monitoring study. *Psychol Addict Behav, 28*(1), 62-73. doi:10.1037/a0033588

Kim, E., & Yoon, M. (2011a). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling-A Multidisciplinary Journal, 18*(2), 212-228.

Kim, E. S., Joo, S.-H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement Invariance Testing Across Between-Level Latent Classes Using Multilevel Factor Mixture Modeling.

Structural Equation Modeling: A Multidisciplinary Journal, 23(6), 870-887. doi:10.1080/10705511.2016.1196108

Kim, E. S., Kwok, O.-m., & Yoon, M. (2012). Testing Factorial Invariance in Multilevel Data: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal, 19*(2), 250-267. doi:10.1080/10705511.2012.659623

Kim, E. S., & Willson, V. L. (2014). Measurement Invariance Across Groups in Latent Growth Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(3), 408-424. doi:10.1080/10705511.2014.915374

Kim, E. S., & Yoon, M. (2011b). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling-a Multidisciplinary Journal, 18*(2), 212-228.

Kim, S. Y., Mun, E. Y., & Smith, S. (2014). Using mixture models with known class membership to address incomplete covariance structures in multiple-group growth models. *Br J Math Stat Psychol, 67*(1), 94-116. doi:10.1111/bmsp.12008

Kim, Y., & Muthén, B. O. (2009). Two-Part Factor Mixture Modeling: Application to an Aggressive Behavior Measurement Instrument. *Struct Equ Modeling, 16*(4), 602-624. doi:10.1080/10705510903203516

Lai, K. (2020). Fit Difference Between Nonnested Models Given Categorical Data: Measures and Estimation. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 99-120. doi:10.1080/10705511.2020.1763802

Lai, M. H. C., & Yoon, M. (2014). A Modified Comparative Fit Index for Factorial Invariance Studies. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 236-248. doi:10.1080/10705511.2014.935928

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics, 34*(1), 1-14.

Little, T. D. (1997). Mean and covariance structures(MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*(2), 53-76.

Liu, L., Ma, J. Z., & Johnson, B. A. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Stat Med*, *27*(18), 3528-3539. doi:10.1002/sim.3205

Liu, L., Strawderman, R. L., Cowen, M. E., & Shih, Y. C. (2010). A flexible two-part random effects model for correlated medical costs. *J Health Econ*, *29*(1), 110-123. doi:10.1016/j.jhealeco.2009.11.010

Liu, Y., Millsap, R. E., West, S. G., Tein, J. Y., Tanaka, R., & Grimm, K. J. (2016). Testing Measurement Invariance in Longitudinal Data With Ordered-Categorical Measures. *Psychol Methods*. doi:10.1037/met0000075

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement, 33*(4), 251-265.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Charlotte, N.C.: Information Age Publishing.

Lubke, G. H., & Muthen, B. O. (2005). Investigating Population Heterogeneity With Factor Mixture Models. *Psychological Methods, 10*, 21-39. doi:10.1037/1082-989X.10.1.21.supp

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample Size in Factor Analysis: The Role of Model Error. *Multivariate Behav Res, 36*(4), 611-637. doi:10.1207/S15327906MBR3604_06

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84-99.

Mackinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behav Res, 39*(1), 99. doi:10.1207/s15327906mbr3901_4

McTernan, M., & Blozis, S. A. (2014). Longitudinal Models for Ordinal Data With Many Zeros and Varying Numbers of Response Categories. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 216-226. doi:10.1080/10705511.2014.935915

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural equation modeling, 14*(4), 611-635.

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

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

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

Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Muthén, B. (2001). *Two-part growth mixture modeling*. [Unpublished Manuscript]. (Article_094).

Muthén, B. O. (1989). Tobit factor analysis. *British Journal of Mathematical and Statistical Psychology, 42*, 241–250.

Muthén, L., & Muthén, B. O. (1998–2017). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nansel, T. R., Overpeck, M., Pilla, R. S., Ruan, W. J., Simons-Morton, B., & Scheidt, P. (2001). Bullying behaviors among US youth: Prevalence and association with psychosocial adjustment. *JAMA, 285*(16), 2094-2100.

Nevitt, J., & Hancock, G. R. (2004). Evaluating Small Sample Approaches for Model Test Statistics in Structural Equation Modeling. *Multivariate Behavioral Research, 39*(3), 439-478. doi:10.1207/s15327906mbr3903_3

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 535-569. doi:10.1080/10705510701575396

Olsen, M. K., & Schafer, J. L. (2001). A Two-Part Random-Effects Model for Semicontinuous Longitudinal Data. *Journal of the American Statistical Association, 96*(454), 730-745. doi:10.1198/016214501753168389

Rhemtulla, M. B.-L., Patricia É.; Savalei, Victoria. (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). doi:10.1037/a0029315.supp

Ryu, E. (2014). Factorial invariance in multilevel confirmatory factor analysis. *British Journal of Mathematical and Statistical Psychology, 67*(1), 172-194.

Satorra, A. (1990). Robustness issues in structural equation modeling: A review of recent developments. *Quality and Quantity, 24*(4), 367-386.

Satorra, A., & Bentler, P. M. (2010). Ensuring Positiveness of the Scaled Difference Chi-square Test Statistic. *Psychometrika, 75*(2), 243-248. doi:10.1007/s11336-009-9135-y

Savalei, V. (2014). Understanding Robust Corrections in Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 149-160. doi:10.1080/10705511.2013.824793

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 161-164.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika, 52,* 333–343.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5*(2), 230-240.

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

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

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. E., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research, 25*(1), 78-90.

Steiger, J. H. (1989). *EzPATH: Causal modeling*. Evanston, IL: SYSTAT.

Steiger, J. H. (2009). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(4), 411-419. doi:10.1080/10705519809540115

van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J., & Muthen, B. (2013). Facing off with Scylla and Charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Frontiers in Psychology, 4*, 15. doi:10.3389/fpsyg.2013.00770

Vandaele, W. (1981). Wald, Likelihood ratio, and Lagrange multiplier tests as an F test. *Economics Letters, 8*, 361-365.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-70. doi:10.1177/109442810031002

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*(1), 1-27.

Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *Plos One, 10*(7), e0129606. doi:10.1371/journal.pone.0129606

Yoon, M., & Millsap, R. E. (2007). Detecting Violations of Factorial Invariance Using Data-Based Specification Searches: A Monte Carlo Study. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 435-463. doi:10.1080/10705510701301677

You, S., Kim, E., & Kim, M. (2014). An Ecological Approach to Bullying in Korean Adolescents. *Journal of Pacific Rim Psychology, 8*(01), 1-10. doi:10.1017/prp.2014.1

Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*(1), 165-200.

Yuan, K. H., Chan, W., & Bentler, P. M. (2000). Robust transformation with applications to structural equation modelling. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 31-50.