UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

RESOLVING THREE IMPORTANT ISSUES ON MEASUREMENT INVARIANCE USING BAYESIAN STRUCTURAL EQUATION MODELING (BSEM)

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

DEXIN SHI
Norman, Oklahoma
2016

RESOLVING THREE IMPORTANT ISSUES ON MEASUREMENT INVARIANCE USING BAYESIAN STRUCTURAL EQUATION MODELING (BSEM)

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Hairong Song, Chair

Dr. Robert Terry, Co-Chair

Dr. Taehun Lee

Dr. Lori Snyder

Dr. David Bard

Dr. Kevin Grasse

For my parents and friends…

In loving memory of

my grandfather(s), ZHENGMING OU & ZHIZHANG SHI

and My uncle DATONG OU

Acknowledgements

I realize that “acknowledgements” is actually the most difficult part to write in my entire dissertation. I have been amazingly fortunate to have many wonderful people to support my Ph. D. study; and I have first ever felt my writing skill is so adequate when I try to express my deepest gratitude to all of them.

First, I am grateful to have two greatest advisors in graduate school. Dr. Hairong Song offered me an opportunity to step into the fascinating field of quantitative psychology in 2010. Her continuous support and encouragement helped me overcome many challenges and crisis situations in my six years’ study. Dr. Robert Terry provided me opportunities to work on multiple projects, and also gave me freedom to explore my own research interests. Whenever I need help or felt lost, Robert is always “next door” to give advice and guidance. Working with my advisors motivated me to pursue a future career in academia; and I hope that one day I could be a “great advisor” to my students, just like Hairong and Robert have been to me.

Besides my advisors, I would like to thank other faculty members in the quantitative psychology program. Dr. Taehun Lee has been a great collaborator and role model for me in academic life. His wisdom and knowledge always inspired me. Dr. Joe Rodgers is one of the greatest teachers that I have ever had; I learnt a lot from him and I am grateful for his help during the SMEP conference at Vanderbilt. Dr. Jorge Mendoza kept providing me with help as the department chair. I am also grateful to him for serving on my thesis committee, and giving me suggestions/information on job seeking.

In addition, I would like to thank my advisory committee members, Dr. Lori Snyder, Dr. David Bard and Dr. Kevin Grasse for their insightful suggestions and

comments on my General Exam and dissertation. I am also indebted to Dr. Lori Snyder for her support on several research projects, and other help she provided me as the graduate liaison.

Moreover, my sincere thanks goes to Dr. Scott Wilson, Dr. Leslie Williams, and Dr. Nicole Campbell for their supervision and guidance when I worked as research assistant at the K20 Center. I am also grateful to the following staff members at University of Oklahoma, for their various forms of support and patient help during my study—Ms. Victoria Willis, Ms. Kathryn Paine, Ms. Abby Young, Mr. Don Rodgers, Ms. Emilie Gordon, Mr. Kamran Sadeghi, and Mr. Lewis Sarkey.

Last but not least, I would like to thank my family and friends. Nothing can express my love for them.

Table of Contents

List of Tables

List of Figures

Abstract

Measurement invariance concerns whether the constructs' measurement properties (i.e., relations between the latent constructs of interest and their observed variables) are the same under different conditions. Without establishing evidence of measurement invariance, corresponding cross-condition comparisons are questionable. Although different systems have been developed in conducting measurement invariance tests, a few important issues shared by those systems remain unsolved. The current dissertation tries to use Bayesian Structural Equation Modeling (BSEM) to address three major imperative issues in studying measurement invariance. First, a new, reliable measure is developed to select a proper (i.e. truly invariant) reference indicator. Second, the issue of locating non-invariant parameters is addressed by using the Bayesian Credible Interval (BCI). Third, posterior distribution is employed to evaluate empirical consequences of non-invariance; specifically, the aim is to interpret non-invariance in terms of expected differences in observed scores across levels of latent trait (or expected differences in latent trait conditioning on observed test scores), and to provide relevant confidence limits. A series of simulation analyses show that the proposed method performs well under a variety of data conditions. An empirical example is also provided to demonstrate the specific procedures to implement the proposed methods in applied research. Extensions and limitations are also pointed out.

Chapter 1: Introduction

Measurement invariance concerns whether the constructs' measurement properties (i.e., relations between the latent constructs of interest and their observed variables) are the same under different conditions. Mellenbergh (1989) gave a formal definition of measurement invariance in terms of conditional probability distributions of the observed scores as shown in Eq. (1).

$$f(X \mid W, V) = f(X \mid W) \quad (1)$$

In general, measurement invariance means conditional independence of observed scores (*X*) given the underlying latent variable W, regardless of any different measurement conditions (*V*) of interest. In practice, these conditions could include different subgroups of a population (e.g., national cultures, ethnical groups, genders), occasions of measurement (e.g., repeated measure data), and different test settings (e.g., paper-pencil vs. web-based test) (Meade & Wright, 2012).

Without establishing measurement invariance, any observed differences across conditions may be simply a reflection of differences in the psychometrical properties of the measures under use, but not the actual differences in the constructs that researchers are desired to test. In this sense, measurement invariance is a very important issue to consider when a measure is applied across different conditions. For example, measurement invariance has been recognized as a prerequisite for examining mean differences across groups or mean changes over time. When invariance of factor loadings and intercepts holds across groups, subjects (from different groups) with the same levels of a latent construct have the same expected observed scores on the measures (Drasgow & Kanfer, 1985). Under this condition the cross-group difference in observed means can be unambiguously interpreted as the true mean difference of the underlying construct. Otherwise, if measurement invariance is not tenable, the observed mean difference could be contaminated by the difference in the psychometric properties of the measures being used.

Among the existing statistical approaches to testing measurement invariance, multiple-group confirmatory factor analytic (CFA; Jöreskog, 1971; McGaw & Jöreskog, 1971) approach has been widely used, with measurement invariance being tested from the perspective of factorial invariance. In general, assuming the observed variables are multivariate normally distributed, factorial invariance has implication for measurement invariance. A technical discussion of measurement invariance and factorial invariance can be referred to Meredith (1993).

Over years, different systems in the framework of CFA have been developed to conduct factorial invariance tests. They differ in labels for different levels of invariance, order and procedures of the series of tests, etc. (see Vandenberg & Lance, 2000 for a comprehensive review). However, they share a few imperative issues that still remain unsolved (Millsap, 2005; Millsap & Meredith, 2007). Of the interest in this dissertation are three major imperative issues, including 1) how to select proper reference indicators; 2) how to locate specific non-invariant parameters; and 3) how to evaluate the consequences of non-invariance.

Those three issues perplex methodologists for many years, and directly obstruct the soundness of measurement invariance studies. First, in using multiple-group CFA techniques to test for factorial invariance, a common method of model identification is to set the factor loading (and intercept as well for model with mean structure) of a particular item to be equal across groups (Reise, Widaman, & Pugh, 1993). The item chosen for this purpose is referred to as a reference indicator (RI). The latent factor is therefore scaled by the RI, and other factor parameters are then estimated in reference to the metric of the RI (Cheung & Rensvold, 1999; Johnson, Meade, & DuVernet, 2009; Meade & Wright, 2012). However, "This creates a dilemma. The reason one wishes to estimate the constrained model in the first place is to test for factorial invariance, yet the procedure requires an a priori assumption of invariance with respect to the referents." (Rensvold & Cheung, 1998). Selection of RIs has been shown to be critical in detecting invariance or non-invariance. When an inappropriate item is chosen to be a RI, severe Type I or Type II errors are expected in testing factorial invariance (e.g., Johnson, Meade, & DuVernet, 2009). It is obvious that how to appropriately select a RI determines whether the true status of invariance could be detected using the multiple-group CFA method.

Second, Within the framework of multiple-group CFA, testing for factorial invariance involves fitting a series of models with stronger forms of equality constraints increasingly imposed.[1] One can determine the tenability of a specific equality constraint by testing the significance of chi-square difference between the two nested models, one with the equality constraint imposed and the other without those constraints. If the test turns out to be non-significant, one can conclude no cross-group differences on the tested parameters. If the test is significant, cross-group difference exists in at least one of the parameters. When some but not all parameters are found invariant across groups, partial invariance is said to occur (Byrne & Shavelson, 1989; Widaman & Reise, 1997). Then one would need to locate specific unequal parameters if they decide to allow those parameters to be freely estimated in the subsequent multiple-group analyses (Widaman & Reise, 1997). In fact, an increasing amount of empirical studies have provided provision for partial invariance (see Schmitt & Kuljanin, 2008; Vanderberg & Lance, 2000). Thus, how to locate nonequivalent parameters apparently becomes imperative in presence of partial invariance.

Third, mainstream approaches for testing non-invariance rely on statistical significant tests, by which very tiny non-invariance is likely to be detected as sample size increases, even though it makes negligible practical influence. Thus, researchers have suggested that non-invariance should be understood as a “continuum”, rather than a dichotomous reject/not reject decision (Nye & Drasgow, 2011). Therefore, it is meaningful to differentiate between statistically significant non-invariance and practically significant non-invariance, especially from the applied users’ perspective. However, currently there is little literature focusing on evaluating empirical consequences of non-invariance.

In summary, despite of their importance, the three (abovementioned) issues still remain under-addressed in both empirical and methodological work. The goal of this dissertation is to address these issues by using Bayesian approach in the context of multiple-group CFA. In the rest of this dissertation, I first briefly introduce multiple-group CFA model and factorial invariance (Section 1.1). Then I focus on detailed discussions on the three issues, the current methods of dealing with each issue, and the pros and cons of each of the methods (Sections 1.2-1.4).

In Chapter 2, I propose the Bayesian approach for solving the three abovementioned issues. The performances of the proposed BSEM methods are evaluated and compared with existing approaches using three Monte Carlo simulation studies in Chapter 3. Specifically, in study I (Section 3.1), I focus on selecting proper (i.e. truly invariant) reference indicator by extending the approach of using informative priors with small-variance and zero-mean. In study II (Section 3.2), the issue of locating non-invariant parameters is addressed by using the Bayesian Credible Interval (BCI). With this approach, the non-invariant parameters can be located by fitting a single model, instead of running a series of models as other approaches do. In study III (Section 3.3), I investigate the usage of the Bayesian posterior distribution to evaluate empirical effect of non-invariance. The non-invariance is interpreted in terms of the expected differences in observed scores across levels of latent trait, as well as the expected differences in latent trait conditioning on the observed test (sum) scores. Different from the existing approaches, the Bayesian method could incorporate the information of sampling errors and provide relevant confident limits.

In Chapter 4, an empirical example is provided to demonstrate the uses of the proposed BSEM methods with real data. Finally, in Chapter 5, I discuss the implications and possible extensions of the proposed BSEM methods. Limitations and future directions are also pointed out.

1.1 Factorial Invariance and the Tests

Confirmatory factor analytic models (CFA; Jöreskog, 1971; McGaw & Jöreskog, 1971) have been widely used to test for factorial invariance across groups or measurement occasions in past decades (refer to Millsap & Meredith (2007) for a thorough discussion on the historical issues of factorial invariance). A standard multiple-group CFA model states the linear relationship between observed variables and latent factors for multiple groups simultaneously. The model can be expressed as

$$\mathbf{y}^{(j)} = \boldsymbol{\tau}^{(j)} + \boldsymbol{\lambda}^{(j)}\boldsymbol{\xi}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \qquad (2)$$

where *j* represents group membership for the vector of observed variables of $\mathbf{y}$, implying that all parameters in the model can differ across groups, $\boldsymbol{\tau}$ represents the intercept vector, $\boldsymbol{\lambda}$ denotes the factor loading matrix, $\boldsymbol{\xi}$ is the latent score matrix, and $\boldsymbol{\varepsilon}$ represents the unique factor vector. Implied by the model, the means and covariances of the observed variables can be expressed, respectively, in matrix forms as:

$$\boldsymbol{\mu}^{(j)} = \boldsymbol{\Lambda}^{(j)}\boldsymbol{\alpha}^{(j)} + \boldsymbol{\tau}^{(j)} \qquad (3)$$

$$\boldsymbol{\Sigma}^{(j)} = \boldsymbol{\Lambda}^{(j)}\boldsymbol{\Phi}^{(j)}\boldsymbol{\Lambda}'^{(j)} + \boldsymbol{\Theta}^{(j)} \qquad (4)$$

where $\boldsymbol{\mu}$ is a vector of population means of the observed variables, $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{\alpha}$ is the vector of the latent factor means, $\boldsymbol{\tau}$ is a vector of intercepts, $\boldsymbol{\Sigma}$ is the population variance-covariance matrix for the observed variables, $\boldsymbol{\Phi}$ is the covariance matrix for latent factors, and $\boldsymbol{\Theta}$ is a variance-covariance matrix among the residuals.

Many forms or levels of factorial invariance have been proposed in the literature (e.g., Byrne, Shavelson, & Muthén, 1989; Horn, McArdle, & Mason, 1983; Jöreskog, 1971; Meredith, 1993; Steenkamp & Baumgartner, 1998; Widaman & Reise, 1997). The two general categories of factorial invariance include configural invariance and metric invariance. Configural invariance is met when same factor structure (i.e. same number of factors and same salient factor pattern) is found across different conditions. Within metric invariance are three commonly-used sublevels, including weak factorial invariance, strong factorial invariance, and strict factorial invariance, with an increasing level of equality constrains added to each latter level of invariance. Specifically, weak factorial invariance holds when factor loadings ($\mathbf{\Lambda}$) are found to be equal across conditions; strong invariance holds when factor loadings ($\mathbf{\Lambda}$) and intercepts ($\boldsymbol{\tau}$) are both found to be equal, which implies that the observed mean differences ($\boldsymbol{\mu}$) are a reflection of true mean differences ($\boldsymbol{\alpha}$) on latent variables (see Equation 3); strict factorial invariance holds when factor loadings ($\mathbf{\Lambda}$), intercepts ($\boldsymbol{\tau}$), and unique variance ($\mathbf{\Theta}$) are all equal across conditions, which implies that the differences in observed means ($\boldsymbol{\mu}$) and variances-covariances ($\mathbf{\Sigma}$) reflect true differences in means ($\boldsymbol{\alpha}$) and variances of the latent variables ($\mathbf{\Phi}$), respectively (see Equation 4). Strict factorial invariance is regarded as a necessary condition of measurement invariance; however, it is also considered as too restrictive to be met in reality.

In general, testing for factorial invariance involves a series of likelihood ratio tests (LRT, Kim & Yoon, 2011; Kim & Cohen, 1995). It begins with fitting a baseline model using one of the two available approaches: free baseline approach or constrained baseline approach. The free baseline approach allows all parameters being freely estimated except for those constrained for model identification purpose. A well-fitted baseline model supports configural invariance. Then metric factorial invariance tests are conducted in order of weak invariance, strong invariance, and strict invariance by adding the associated equality constrains. For example, in a test of weak invariance, one sets all factor loadings to be equal across conditions and then evaluates the significance of chi-square difference between this model and the baseline model using the likelihood ratio test. If the test indicates non-significance, which means the model with constrains fits the data as well as the baseline model, researchers would accept the constrained model and continue to add further constrains to test a higher level of invariance. If the hypothesis of full weak invariance is not retainable, researchers may adopt partial invariance methodology, by which to locate the non-invariant indictor variables and to freely estimate their parameters. Detailed discussions on partial invariance are offered in a later section of this dissertation.

In contrast to the free baseline approach, constrained baseline approach fits the baseline model by imposing full equality constrains on all model parameters; that is, it begins with fitting a strict factorial invariance model as the baseline. Then weaker forms of invariance are tested by increasingly relaxing those constrains and the chi-square differences between nested models are evaluated using LRT. The constrained baseline approach is used and recommended by some researchers (Kim & Yoon, 2011). However, one problem of the constrained baseline approach is that the non-invariant parameters (if present) are also fixed to be equal as baseline; therefore, the constrained baseline model may not produce a reasonably well fit to data. As noted by Maydeu-Olivares and Cai (2006), a well-fitted baseline model is required as a statistical premise for the LRT; only under conditions which the baseline model fits well does the difference between the nested models follow a central chi-square distribution under the null hypothesis. In this sense, the free baseline approach is more appropriate from a statistical perspective. By comparing the two approaches, researchers also found that using the constrained baseline approach to test factorial invariance may produce unreliable conclusions, such as inflation of type I error rates (see Stark, Chernyshenko & Drasgow, 2006; Kim & Yoon, 2011 for more comparisons between these two approaches). In this dissertation, the goal is to address questions related to testing for factorial invariance using the free baseline approach, namely, selecting referent indicators and locating non-invariant parameters.

1.2 Selection of Reference Indicators

Selection of RIs has been recognized as an important issue in the literature of factorial invariance. It is well known that in fitting a single-group CFA, the metric of latent variables needs to be set to make the model identified. One can either set the factor variance to unity or assign one of the factor loadings to unity (Bollen, 1989). When fitting a multiple-group CFA, there are multiple ways to identify the multiple group models, and these different methods are equivalent in terms of goodness of fit.[2] However, for the purpose of studying factorial invariance, scaling factor variances to unity for both groups is not applicable. If the data do not support this equality constrain, the estimation of model parameters may be biased (Rensvold & Cheung, 1998) and the true differences in between-group factor variances may be shifted to observed differences in between-group factor loadings. Therefore, this model identification method is not recommended to use for testing factorial invariance (Yoon & Millsap, 2007). Alternatively, one commonly-used method for multiple-group CFA identification is to use reference indicators (RI).Specifically, one can select one arbitrary group as the reference group and fix its factor variance to one. In addition, the factor loadings for the selected RI are constrained to be equal across groups. In doing this, there is only one set of estimated coefficients that reproduces the data optimally. In other words, the multiple-group model is identified. In the meanwhile, since other parameters are estimated in reference to the factor variance in group one and the selected RI, the scale of the multiple-multiple group model is set so that the corresponding parameters are comparable across groups (Cheung & Rensvold, 1999; Johnson, Meade, & DuVernet, 2009; Meade & Wright, 2012).

Research has shown that if RIs are not truly invariant, factorial invariance tests can be jeopardized and the true state of invariance could be greatly obscured. For example, when an item that is not metrically invariant is inadvertently selected as a RI, truly invariant items could be erroneously detected as non-invariant items and truly non-invariant items could be erroneously detected as invariant (Johnson, Meade, & DuVernet, 2009). Thus, severe Type I or Type II errors are expected in factorial invariance tests when an inappropriate item is selected as a RI. Similarly, empirical analysis has shown that in fitting second-order growth curve models, choice of RIs could also have substantial influences on both model fit and estimates of growth trajectories, when full factorial invariance is not tenable (e.g., Ferrer, Balluerka, & Widaman , 2008; Widaman et al., 2010).

Researchers have proposed different methods to deal with issues associated with selecting RIs. Cheung and Rensvold (1998) proposed the idea that instead of using one fixed item as a RI, each single item can serve as a RI in turn. At first, all possible non-ordered pairs are generated by taking two items at a time without repetition. So for a measure with *n* items, one needs to generate a total of *n* (*n*-1)/2 such item pairs. Then a so-called factor-ratio test is conducted to identify invariant items; that is, within each pair, two models are fitted with either item as the RI to test the invariance of the other. A significant factor-ratio test suggests that at least one item within the tested pair is non-equivalent across groups; otherwise, this pair of items is considered invariant. After identifying all the non-invariant and invariant pairs, a stepwise partitioning procedure is then used to screen out each item as invariant or non-invariant. Taking a four-item test as an example, if items (1, 2), (1, 3), and (1, 4) are non-invariant pairs, whereas rest of the pairs are invariant (i.e. items 2, 3; 2, 4; and 3, 4), then items 2, 3 and 4 are concluded as the final set of invariant items.

Cheung and Rensvold (1998)'s approach allows researchers to detect non-invariance without using a specific item as the referent. Therefore, it reduces the possible negative influences associated with using non-invariant items as RIs. The utility of this approach was further supported by research using simulated data (French & Finch, 2008). However, several disadvantages have prevented this approach from being widely used in real world research. First, the implementation of this approach requires fitting two models for each of the n (n-1)/2 item pairs, which could be labor intensive, especially under conditions with large number of items (French & Finch, 2008; Yoon & Millsap, 2007; Cheung & Lau, 2012). Moreover, it is possible that no single set of invariant items can be determined conclusively with this method, as demonstrated by French and Finch (2008). In practice, any Type I or Type II error occurring in the factor-ratio tests could cause this indeterminate conclusion. For example, for a four-item test, if non-invariant pairs include items (1, 2), (1, 3), and (2, 4), whereas the invariant pairs include items (1, 4) and (2, 3), then no single set of invariant items can be determined conclusively.

Later, Cheung and Lau (2012) proposed a new method aiming to overcome the drawbacks of Cheung and Rensvold (1998) procedure. Using this approach one needs to first create a new parameter that represents the between-group difference for each tested parameter, and then the bias-corrected bootstrap confidence interval is generated for each of such differences. The multiple-group CFA model is identified by using an arbitrary item as the RI. For example, let λ_{i1} and λ_{i2} represent the factor loadings of the *i*th item for group 1 and group 2, respectively. Suppose researchers use item 1 as RI by imposing such constrains: $\lambda_{11} = \lambda_{12} =1$. In order to test invariance for the pairs containing items 1 and 2, a new parameter T_{12} needs to be created. T_{12} can be simply expressed as $T_{12} = \lambda_{21} - \lambda_{22}$, where λ_{21} and λ_{22} are factor loadings of item 2 estimated using item 1 as RI. If the 95% bootstrap confidence interval does not include zero, the pair of items (1, 2) is concluded as non-invariant. For pairs not including item 1, the corresponding parameters can be obtained using existing parameters analytically. For instance, $T_{23,}$ which is used to test invariance for pairs containing item 2 and item 3, can be expressed as: $T_{23} = \lambda_{31}/\lambda_{21} - \lambda_{32}/\lambda_{22}$ (refer to Cheung & Lau, 2012 for a detailed technical rationale).

The major advantage of Cheung and Lau (2012)'s approach is that it allows the invariant and non-invariant item pairs to be detected by fitting a single multiple-group CFA model. Therefore, it greatly reduces the workload, compared to the Cheung and Rensvold (1998) approach, and can be conveniently generalized to test invariance for intercepts and latent means/variances. Nevertheless, the stepwise partitioning procedure is still required to screen out the set of invariant items. Therefore, it is subject to the same issue as the Cheung and Rensvold (1998) method; that is, identifying a conclusive set of invariant items is not guaranteed. In addition, the bias-corrected bootstrap confidence interval may not perform sufficiently well under some conditions, such as small samples, as pointed out by Cheung and Lau (2012).

Constrained baseline approach is an alternative to selecting RIs (Stark, Chernyshenko & Drasgow, 2006; Kim & Yoon, 2011). In this approach, all items are constrained to be equivalent across groups in the baseline model. Then equality constraints are relaxed for the single tested item. The difference in the model fit between the two models is used to evaluate invariance of the tested item. This procedure is repeated for all of the other items. An item chosen to be RI is the one that yields non-significant cross-group differences but has the largest estimated factor loading. The constrained baseline approach implicitly uses all (other) items as RIs to test cross-group differences in items. The idea has been widely adopted in research of item response theory (IRT; Meade & Wright, 2012). In multiple-group CFA analysis, however, research has shown that when the percentage of non-invariant items is large, this approach can yield inflated Type I error rate, that is, the chance of identifying truly invariant items as non-invariant is enhanced (Kim & Yoon, 2011).

1.3 Locating Non-invariant Parameters

Rejecting a null hypothesis of full invariance at any given level typically does not provide direct information on which specific parameters differ across groups. When factorial invariance does not hold, locating specific non-invariant parameters becomes necessary for at least two reasons. First, by knowing which parameters are non-invariant, researchers have an opportunity to explore potential causes of the detected non-invariance, which can be substantively meaningful in applied research. For example, unequal factor loadings implies that the association between the construct of interest and non-invariant items may be weaker (or stronger) in one group than the others being compared. The researchers may then be able to identify whether the lack of invariance on loadings is due to inappropriate translation (e.g. in cross-cultural studies) or different understanding of item contents across groups. If the non-invariance occurs in some of the measurement intercepts, the origin of items may be different across groups, and such non-invariance may be attributed to factors such as social desirability, usage of different reference frameworks, etc. (Chen, 2008).

Secondly, locating unequal parameters is closely related to the practice of fitting models with partial measurement invariance. When full factorial invariance is not tenable, it is suggested that fitting models with partial invariance is useful in various empirical modeling settings. For example, studies have suggested that when testing for possible differences in latent means, latent variances, or structural relations with other constructs, fitting models with non-invariant parameters freely estimated would produce more accurate parameter estimates than the model with all parameters being constrained to equality (Shi, Song & Lewis, 2016; Muthén & Asparouhov, 2013). Using simulated data, Liao et al (2015) found that when fitting second-order latent growth curve models, if full factorial invariance was not fulfilled, models with partial invariance yielded less biased estimates than models with full invariance specified.

In practice, locating non-invariant parameters is typically guided by model modification index (MI) in maximum-likelihood estimation. Each MI can be understood as a one-degree-of-freedom likelihood ratio test (LRT) performed between the models with and without a certain constraint. In the context of testing factorial invariance, rejecting a hypothesis of full invariance implies the presence of improper equality constraint(s) across groups. MI can be used to identify where each improper equality constraint is located in the model. For example, if the model with all factor loadings set to be equal is not supported, one can first free the equality constraint imposed on the loading associated with the largest MI value. This procedure can be repeated until the model fit becomes acceptable, and there are no additional noticeably large MIs. Researchers have been concerned about the uses of MI in locating non-invariant parameters (e,g., MacCallum, Roznowski & Neocowitz,1992) and have found that MI does not perform efficiently in locating non-invariant parameters under such conditions as small sample size, small magnitude of non-invariance, and large number of non-invariant variables (Yoon & Millsap, 2007). In addition, since MI values change unceasingly at any time when a constraint is relaxed, using MI to locate non-invariant parameters typically requires fitting a series of models in many popular SEM software (e.g. Mplus), which could become tedious in some cases.

1.4 Evaluating the Consequences of Non-Invariance

Currently, the likelihood ratio test is the mainstream approach for testing non-invariance. As discussed earlier, the likelihood ratio tests involve comparing two (or a series of) nested models which have different levels of constrains. The statistical decisions are made by the chi-square tests; if significant, the less constrained model is conclude to fit the data significantly better, and thus the hypothesis of factorial invariance would be rejected. Nevertheless, the usage of chi-square based tests has been questioned by methodologists. One major criticism on the usage of chi-square statistics is that the chi-square test is sensitive to sample size (*N*). As sample size (i.e. *N-1*) is a "multiplier" of the chi-square variate; with sufficiently large sample, even small differences between the hypothesized model and data would be statistically significant; and power to detect any trivial difference would approach to one as sample size increases (Marsh, Hau & Grayson, 2005; West, Taylor, & Wu, 2012; Bentler & Bonett, 1980). As a result, even if the non-invariance is minor and makes negligible practical influence, the hypothesis of factorial invariance is still very likely to be rejected as sample size increases. For example, as demonstrated by Meade (2010) in the context of Item Response Theory (IRT) models, even trivially small non-invariance often leads to statistically significant results with sample sizes reach 1,000 per group. Therefore, from the applied users' perspective, in addition to testing the existence of non-invariance, it is also meaningful to considering the practical effect size of the detected non-invariance.

Currently, there is little literature focusing on evaluating empirical consequences of non-invariance. One possible approach to evaluate the empirical consequences of non-invariance is to examine the purpose of the measure in use, and thus translating the non-invariance into practical outcomes. For example, Millsap and Kwok (2004) evaluated the impact of partial invariance on selection based on composite scores for two populations. The simulation study showed that departures from invariance weakened the accuracy of selection.

In addition, researchers also proposed several effect size indices for the purpose of evaluating and comparing the magnitude of non-invariance. Within the framework of SEM, Nye and Drasgow (2011) proposed an item-level effect size measure for measurement non-invariance (d_{MACA}), which is defined as the following.

$$d_{MACS} = \frac{1}{SD_{i\,p}} \sqrt{\int (\hat{X}_{iR} - \hat{X}_{iF} \mid \xi)^2 f_F(\xi) d\xi} \qquad (5)$$

$\hat{X}_{iR}$ and $\hat{X}_{iF}$ represent the expected observed responses to item *i* with latent score ξ for the focal reference group and focal groups, respectively. $f_F(\xi)$ indicates the distribution of the latent traits for the focal group. SD$_{\mathrm{ip}}$ represent the pooled standard deviation for item *i* across the reference and focal groups, given by

$$SDip = \frac{(N_P - 1)SD_R + (N_F - 1)SD_F}{(N_R - 1) + (N_F - 1)} \text{.} \qquad (6)$$

For both reference and focal groups, the expected overserved responses can be regressed on the latent factor scores as a linear function (see Equation 2) The d_{MACA} measure can be roughly represent the area between the two regression lines, which express the “overall” amount of non-invariance across the domain of latent trait. In addition, after taking the pooled standard deviation into formulation, d_{MACA} can be interpreted in the standardized metric and the magnitudes can be directly compared across different items and studies.

Chapter 2: Bayesian Structural Equation Modeling (BSEM) Approach

2.1 Bayesian Structural Equation Modeling (BSEM)

The differences between the traditional statistical methods with the frequentist view and the Bayesian methods have been discussed in many literatures (Brooks, 2003; Dienes, 2011; Kruschke, Aguinis & Joo, 2012). The traditional statistical methods rely on null hypothesis significance testing. Specifically, the frequentist statistics views the parameters as constants, and the significance tests for tested parameters focus on *p* values that is the probability of obtaining the observed data or something more extreme, if the null hypothesis were true. The Bayesian approach, however, aim to directly provide credibility of possible parameter values based on the observed data. The general idea of the Bayesian methods is to treat model parameters as variables, and thereby evaluating posterior distributions that are derived based on the Bayes' Theorem for the parameters.

In recent years, Bayesian approach has been increasingly applied for fitting complex models in behavioral sciences that involve latent variables and many parameters (e.g. Song & Ferrer, 2013; Lee & Song, 2004; Serang et al, 2015). As follows, a general Bayesian estimation procedure is briefly introduced in the context of SEM.

Let **M** be an arbitrary SEM model with the unknown parameters in a vector $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{M})$ be the prior distribution of the parameter, and **Y** represent the observed data. A standard Bayesian approach requires the evaluation of the posterior distribution of $\boldsymbol{\theta}$ given **Y** (i.e., $p\ (\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M})$). This can be obtained by $P(\boldsymbol{\theta}\,|\,\mathbf{Y},\mathbf{M}) \propto P(\mathbf{Y}\,|\,\boldsymbol{\theta},\mathbf{M}) \bullet P(\boldsymbol{\theta}\,|\,\mathbf{M})$ based on the Bayes' Theorem, where $p\ (\mathbf{Y}|\ \boldsymbol{\theta}, \mathbf{M})$ is the likelihood of observing data **Y** conditional on the parameters $\boldsymbol{\theta}$, and $p\ (\boldsymbol{\theta}|\mathbf{M})$ is the prior probability of the parameters $\boldsymbol{\theta}$. Suppose the posterior distribution $p\ (\boldsymbol{\theta}|\mathbf{Y}, \mathbf{M}\ )$ is analytically obtained, numerical integration would be used to obtain the posterior mean and posterior variance for the model parameters. However, when the model involves latent variables and many parameters, the high-dimensional integration often times has no closed form and consequently, the posterior mean and variance cannot be obtained analytically. Markov Chain Monte Carlo (MCMC) method can be used to handle such otherwise intractable calculation. The basic idea is to repeatedly draw random numbers from (full or conditional) posterior distribution and empirically summarize those draws, thereby approximating the mean and variance of the target parameters (Martin, 2005). In Bayesian estimation of the measurement model in SEM, a data augmentation technique is used (see Tanner & Wong, 1987), by which factor scores are treated as unknown parameters and the observed data is "augmented" with factor scores to develop the Bayesian procedure. Ultimately, the posterior distributions of all model parameters could be obtained.

The parameter estimates are then obtained as the empirical means, modes, or medians of the posterior distributions (Song & Lee, 2012). The Bayesian Credible Intervals (BCI) can be obtained based on percentiles of the posterior distribution, and directly interpreted in a probabilistic manner. For example, one can claim that there is a 95% chance that a parameter falls in the 95% BCI, which is generally believed to include the most credible values of the parameter. Therefore, as discussed in Kruschke, Aguinis and Joo (2012), besides serving as a summary of a posterior distribution, the

95% BCI can be used as a tool to decide which parameter values to reject, in analogous to confident interval, although the two differ on a philosophical basis.

To implement MCMC, prior distributions need to be first supplied for the unknown model parameters so that posterior distributions can then be derived by modifying the data likelihood using the priors. Non-informative and informative priors are available. When information about the target population is not known, one typically uses non-informative priors. Examples of non-informative distributions are normal distributions with large variances. Larger variances are associated with larger uncertainty about the parameters. Since such priors carry little or no information about the parameter, the estimation is predominately determined by the data. On the other hand, informative priors refer to useful, prior knowledge of unknown parameters. They often reflect strong substantive beliefs about the parameters, and can be decided based on theory, knowledge of experts, or results from previous studies (Song & Lee, 2012).

Muthén and Asparouhov (2012) recently proposed a new approach to SEM using Bayesian analysis, referred to as BSEM. The basic idea of BSEM is to use informative, small-variance priors to replace parameters fixed at zero under ML, thereby better reflecting substantive theories. Informative priors with small-variance and zero-mean express a relative strong belief that the parameters imposed with such priors are close enough to zero; meanwhile, those parameters are not fixed as exact zeros and therefore, they are still estimable and their significance can then be tested. Estimating those parameters under ML otherwise could lead to non-identified model. This idea of using Bayesian approximate constraints to replace exact constraints showed promising uses in estimating cross-factor loadings and correlated residuals in

CFA (Muthén & Asparouhov, 2012). In this dissertation, I extend the uses of the BSEM approach in selecting reference indicators, locating non-invariant parameters, as well as evaluating the consequences of non-invariance as conducting factorial invariance tests.

2.2 The BSEM Approach for Studying Factorial Invariance

I first define a new parameter, D_{ij}, to represent the cross-group difference [3] in a specific parameter *i* for item *j*. For example, $D_{\lambda 1}$ denotes the metric difference in the factor loadings for item 1 across groups that can be expressed as $D_{\lambda 1} = \lambda_{1\,(1)} - \lambda_{1\,(2)}$, where numbers in the parenthesis represent group membership. For simplicity, the population model is assumed to be a one-factor model with partial invariance in both factor loadings and measurement intercepts. If the factors in the multiple-group CFA model are properly scaled by using one or more truly invariant items as the RIs, the estimates of the factor loadings for the invariant items are expected to be approximately equal across groups. As a consequence, the estimates of $\mathrm{D}_{\lambda \mathrm{j}}$ for the invariant factor loadings are expected to be approximately zero. For those items with unequal factor loadings, however, the estimates of $\mathrm{D}_{\lambda \mathrm{j}}$ should noticeably depart from zero. The same difference parameters can also be defined for other model parameters such as for intercepts ($\mathrm{D}_{\tau \mathrm{j}}$). The first property of the difference measure (D_{ij}) is summarized as follows:

<u>Property 1:</u> *If the latent variables in multiple-group CFA model are identified and scaled by using truly invariant items as the reference indicators, the estimates of Dij are expected to be approximately zero (i.e.* $\hat{D}_{ij} \approx 0$*) for invariant parameters, but noticeably different from zero (i.e.* $|\hat{D}_{ij}| > 0$*) for non-invariant parameters.*

Property 1 suggests that the difference parameter can be used as a valid measure in detecting invariance and non-invariance. However, it holds under a rather restrictive condition, that is, only if truly invariant items are selected as the RIs. What would happen to the difference measure if truly invariant items are unknown, as often the case in real-data analysis? One way to scale the factor in this case is to set factor loadings of all items to be equal across groups. As a result, the metric of the latent factor and other parameter estimates are determined by both invariant and non-invariant items altogether. One would expect that the estimates of D_{ij} are likely to deviate from those that would otherwise be obtained from the model only using invariant items as RIs. Nevertheless, as long as the majority of parameters are invariant, as expected in well-developed instruments, it is reasonable to expect that the estimates of D_{ij} would be closer to zero for invariant parameters than those for non-invariant parameters. Now the relative stances of the estimated D_{ij} for invariant and non-invariant parameters would matter more than the individual estimate of D_{ij} for each parameter. Property 2 can therefore be specified as follows:

<u>Property 2</u>: *If the latent variables in multiple-group CFA model are scaled in the way that the metric of the model parameters can be considered as an good approximation to the metric otherwise set by truly invariant parameter(s) only, the estimates of D_{ij} for invariant parameters should be much closer to zero, compared to those for non-invariant parameters, which can be expressed as*

$$|\hat{D}_{ij}|_{in\text{var}iant} < |\hat{D}_{ij}|_{non-in\text{var}iant}.$$

Compared with non-invariant parameters, truly invariant parameters tend to produce smaller values on $|\hat{D}_{ij}|$ when all items are constrained to be invariant across groups. So it is legitimate to expect that a truly invariant item (i.e., invariant in both the factor loading and the intercept) would be associated with the smallest $|\hat{D}_{ij}|$on both the factor loading and the intercept. I hereby proposed an item level selection index, Δ_j, to quantify a criterion for selecting invariant items that can be used as RIs in the subsequent analysis. The selection index can be expressed as a sum of two standardized difference measures, assuming *Ds* for loadings and intercepts are independent:

$$\hat{\Delta}_j = \frac{|\hat{\mathrm{D}}_{\lambda \mathrm{j}}|}{sd_{\lambda j}} + \frac{|\hat{D}_{\tau j}|}{sd_{\tau j}} \qquad (7)$$

where $\hat{D}_{\lambda j}$ and $\hat{D}_{\tau j}$ are the respective estimates of difference measures on the factor loading and the intercept for item *j*, and $sd_{\lambda j}$ and $sd_{\tau j}$ are the corresponding standard deviations. The item that produces the smallest value of Δ_j is identified as an invariant item.

The standard SEM approach cannot estimate the D_{ij} and the selection index, because it requires fixing both $D_{\lambda j}$ and $D_{\tau j}$ to be exact zero for the purpose of scaling the latent factors. In order to set the metric of the parameters properly and make those D parameters estimable as well, I proposed to impose the Bayesian approximate constraints on the D parameters; that is, instead of constraining those parameters to be zero, I impose informative priors with zero-mean and small-variance for $D_{\lambda j}$ and $D_{\tau j}$. As stated earlier, priors with zero-mean and small-variance ensure the latent factors to be properly scaled; on the other hand, $D_{\lambda j}$ and $D_{\tau j}$ are not fixed and still can be estimated by the Bayesian method. Once $D_{\lambda j}$ and $D_{\tau j}$ are estimated for each item, one can compute the selection index and evaluate its posterior distribution. The item that produces the smallest posterior mean on Δ is considered to have the highest likelihood to be invariant across groups.

After the identified invariant item is set as the RI, the next step is to locate possible non-invariance in other parameters (e.g., factor loadings and intercepts). Property 1 of the difference parameter states that if the scales of the latent variables in the multiple-group CFA model are set by truly invariant items, the estimates of the *D* parameters are expected to be significantly different from zero for non-invariant parameters. The significance of the *D* parameters is determined by examining the 95 percent BCIs. If the interval for a $D_{\lambda j}$ excludes zero, one can conclude that the factor loading for item *j* is not equal across groups. Since one can obtain posterior distributions for all D_{ij} simultaneously, locating non-invariant parameters can be done by fitting a single multiple-group CFA.

Provided that the non-invariant parameters are successfully detected, the consequences of non-invariance can be evaluated in the following two ways. First, the cross-group differences on the observed scores for individuals with the same level of latent trait can be obtained. At the item level, the expected difference on observed responses $E(X_{j(1)} - X_{j(2)} \mid \xi)$, or EDOI can be expressed as[4]

$$EDOI : E(X_{j(1)} - X_{j(2)} \mid \xi) = (\lambda_{j(1)} - \lambda_{j(2)}) \bullet \xi + (\tau_{j(1)} - \tau_{j(2)}) . \qquad (8)$$

where $\lambda_{j(1)}$ and $\lambda_{j(2)}$ represent the factor loadings for item *j* in group 1 and group 2, respectively. $\tau_{j(1)}$ and $\tau_{j(2)}$ are the corresponding intercepts. ξ indicates the latent traits. The expected difference of the observed scores at the test (i.e. total score) level (EDOT) can be expressed in the similar manner as

$$EDOT: E(S_{(1)} - S_{(2)} \mid \xi) = \sum_j (\lambda_{j(1)} - \lambda_{j(2)}) \bullet \xi + \sum_j (\tau_{j(1)} - \tau_{j(2)}) . \qquad (9)$$

$S_{(1)}$ and $S_{(2)}$ represent the test score of the scale (total score across *j* items) for group 1 and group 2, respectively. Other notations are the same as defined previously. Using BSEM, for any given latent score, the two measures defined above can be treated as variables and the corresponding posterior distributions are obtained. In addition to the point estimate, the confident limits can be accessed from the posterior distributions. Therefore, the cross-group differences on observed scores (at both item and test levels) with confident limits can be obtained and plotted across different levels of latent trait.

Secondly, the outcome of non-invariance can also be interpreted in terms of the expected differences on latent traits for individuals with the same test scores (EDLT). That is

$$EDLT: E(\xi_{j(1)} - \xi_{j(2)} \mid S) = \frac{S_{(1)} - \sum_j \tau_{j(1)}}{\sum_j \lambda_{j(1)}} - \frac{S_{(2)} - \sum_j \tau_{j(2)}}{\sum_j \lambda_{j(2)}} . \qquad (10)$$

All notations are the same as previously defined. Using BSEM, the cross-group differences on latent traits are treated as variables with posterior distributions. Consequently, for any given observed test score, the expected cross-group differences on latent traits with corresponding confident limits can be obtained.

In the next chapter, I examine the performance of the proposed BSEM method in selecting invariant items, locating non-invariance, and evaluating the outcomes of non-invariance through simulation studies.

Chapter 3 Simulation Studies

3.1 Study I: The Selection Index in Selecting an Invariant Item as RI

Data Simulation

The data were generated based on a population multiple-group CFA model with continuous indicators. We restrict the number of groups to two and the same number of items loaded on the same single factor in each group. One group serves as the reference group in which the factor mean and factor variance are set to be zero and one, respectively. In the reference group, all factor loadings are simulated with a population value of 0.80, and all intercepts are simulated to be 0 for simplicity. The other group is the focal group in which the population factor mean and factor variance are set to 0.5 and 1.2, respectively. In the focal group, the factor loadings and intercepts are set to be equivalent as the reference group (i.e. factor loadings equal 0.8 and intercepts equal 0), except for the predetermined non-invariant parameters. The population values for the non-invariant factor loadings and intercepts in the focal group are determined according to different simulation scenarios as described below. For both reference and focal groups, the population values for all residual variances are set to 0.36.

Five variables are manipulated in our data simulation: Sample size, number of items, percentage of non-invariant items, source of non-invariance, and magnitude of non-invariance.

Sample size. The two groups are generated with equal number of observations. Sample sizes include 50, 100, 200, and 500 per group. Fifty observations per group are considered as extremely small samples and 500 per group are considered as large in standard SEM literature.

Number of items. We generate the factor models with either five or ten indicators. The numbers are considered to be consistent with the typical lengths for psychological measures in common use.

Percentage of non-invariant items. Two conditions are considered regarding percentages of non-invariant items. For the low contamination condition, 20% of the indicators contain non-invariant parameters. For the high contamination condition, the proportion of non-invariant items is 40%. Those percentages are chosen based on previous simulation studies on testing factorial invariance (French & Finch, 2008; Meade & Wright, 2012).

Source of non-invariance. Non-invariance is simulated either on factor loadings or intercepts, not on both at the same time. All unique variances are simulated to be equivalent across groups.

Magnitude of non-invariance. Under conditions with small cross-group differences, factor loadings in the focal group decrease by 0.2, or intercepts increase by 0.3. Under the large difference conditions, factor loadings in the focal group decrease by 0.4, or intercepts increase by 0.6. The choices for the magnitudes are based on suggestions from previous literature (Kim & Yoon, 2011; Kim, Yoon & Lee, 2012; Meade & Lautenschlager, 2004).

In total, sixty-four (4*2*2*2*2) different scenarios are considered in this study. For each simulation scenario, 500 replications are generated and analyzed with Mplus 7.11(Muthén & Muthén, 1998-2012).

Analysis

In order to select a proper RI, Bayesian multiple-group CFA models are fitted using Mplus 7.11. The generated reference group is intentionally selected as the reference group in which the latent factor was constrained to have zero mean and unit variance. The latent variables are scaled by imposing normal priors of zero-mean and small-variance on cross-group differences (D_{ij}) for all the factor loadings and intercepts. Four different values of variances are considered, including 0.001, 0.01, 0.05, and 0.1. The choices of these prior variances reflect different levels of certainty on the parameter values. For example, if a prior is N (0, 0.001), it indicates the 95% interval of D_{ij} lies between -0.06 and +0.06. Variances of 0.01, 0.05 and 0.1 produce 95% limits of ± 0.20, ± 0.44, and ±0.62, respectively. All other parameters (include the factor mean and factor variance in the focal group, and all residual variances) are freely estimated with non-informative priors imposed. The same models are fitted with 4 different informative priors under all simulated conditions. The final MCMC chain runs for a minimum of 50,000 and a maximum of 100,000 iterations. In order to control for auto-correlation among the MCMC iterations, only every 10th iteration is recorded to describe the posterior distribution. Based on the posterior distributions of the *D*, the selection index (Δ) is then computed for each item. Finally, the item that produced the smallest estimate on Δ is selected as RI. All the above set-ups are applied to every simulated replication. Power rates are computed under each data condition, as the percentage of correctly selecting a truly invariant item as RI, to aid the performance evaluation. The procedure of computing Δ and power rates is automated using SAS 9.3 based on the Mplus outputs (Gagné & Furlow, 2009). Meanwhile, all the simulated data are analyzed using the constrained baseline approach proposed by Rivas, Stark, and Chernshenko (2009).[5] The power rates of selecting RI are compared between this approach and the BSEM under each condition.

<u>Results</u>

The results are summarized in Table 1. In general, the selection index implemented by the BSEM method performs well for most of the simulation scenarios, especially under conditions with low percentage of non-invariance, large magnitude of non-invariance, and large sample size. Specifically, under conditions with 20% of non-invariance, the power of correctly selecting RI is far higher than .90 for almost all cases, regardless of magnitude of non-invariance and sample sizes. Under the high contamination conditions (i.e., 40% of non-invariance), the power rates are still above .80 for most of cases; however, they decrease with decreasing magnitude of non-invariance and decreasing sample sizes. The power rates improve noticeably as the sample size increase. It appears that sample sizes of 100 or above would make the power reach .80 or higher in the case of high contamination conditions. With respect to the number of items, the power rates are higher for the 5- item condition than the 10-item condition when sample size is small. Once the sample size reaches 200 and above, the effect of number of items is no longer detectable. Finally, the choice of prior variances does not significantly affect the power rates when relatively smaller prior variances (i.e., 0.001 and 0.01) are utilized. When larger variances (i.e., 0.05 and 0.1) are used, the power rates are fairly close to those produced by smaller prior variances, except for the conditions where high proportions of small non-invariance are generated for the 10-item case--the power rates are even lower than selecting a correct RI randomly.

The BSEM is found to perform better than the constrained baseline approach, particularly when non-invariance occurs on item intercepts. As shown in Table 1, when percentage or magnitude of non-invariant intercepts is large, the constrained baseline approach fails to select truly-invariant items to be RIs. An extreme case is found when 40% of the intercepts are coupled with large cross-group difference, in which the constrained baseline approach has power of zero to selecting the correct RIs at the sample size of 500. However, this approach seems to outperform the BSEM method when large amount of factor loadings is different across groups at small samples. This may be due in part to the fact the truly invariant items happen (by the simulation design) to have larger factor loadings, compared to those for non-invariant items in our simulated data. In this case, the constrained baseline approach, characterized by selecting an item with non-significant cross-group difference but the largest loading to be RI, would tend to have relatively high power to choose a truly invariant item to be RI.

3.2 Study II: The BSEM Approach in Locating Non-invariance using RI

Data Simulation

The goal for the second set of analysis is to investigate the performance of using the BSEM for locating specific non-invariant parameters, given that an invariant item has already been correctly selected as RI. I consider the same simulation scenarios and thus use the same simulated datasets as in the first set of analysis.

Analysis

One of the truly invariant items is chosen as the RI in this set of analyses. All factor loadings and intercepts of the selected RI are fixed to be equal across groups.

The simulated reference group is still served as the reference group with factor mean fixed to be zero and variance to be one. All other parameters are freely estimated with non-informative priors (i.e., normal priors with large variances). Each MCMC chain runs for a minimum of 50,000 and a maximum of 100,000 iterations, and only every 10th iteration is recorded to eliminate auto-correlations. The non-invariant parameters can be detected by checking the posterior distributions for the Dij. If the 95% BCI for a tested parameter does not include zero, this parameter is identified as non-invariant; otherwise, the parameter is considered to be invariant. So all parameters (except for parameters of the RI) can be tested for invariance simultaneously by fitting a single model.

Type I error and power rates are calculated for performance evaluation. Type I error refers to the cases of incorrectly concluding invariant parameters as non-invariant. In each replication, the Type I error rate is calculated as the number of invariant parameters detected as non-invariant divided by the total number of invariant parameters. Power, on the other hand, represents the probability of correctly detecting non-invariant parameters, which is calculated as the percentage of non-invariant parameters that are correctly detected. For each simulated scenario, power and Type I error rates are reported as the averages across all replications. Such definitions make the results easy to understand and interpret. For example, if the aggregated power rate equals 0.70, we can expect that 70% of the non-invariant parameters are be detected under that specific data condition. Aggregated Type I error rates and power have been used in many previous literature as well (e.g., Meade & Wright, 2012). Since not all replications are successfully analyzed due to non-convergence of the MCMC chains, the admissible solution rates are also computed and reported.

Results

Table 2 shows the admissible solution rates, Type I error rates, and power for the 5-item and 10-item scenarios, respectively. All the models are successfully estimated except for those with small sample sizes of 50, for which the admissible solution rates range from 79% to 88% for the 5-item conditions and from 34% to 44% for the 10-item conditions. The Type I error rates are low across all simulation conditions with a range of 3% to 6%. So in general, as long as the models can be estimated with the Bayesian method, the chance of identifying an invariant parameter as non-invariant is low. The power rates are very similar between the 5-item and 10-item scenarios under the same data conditions. In addition, the proportion of non-invariant parameters seems not to influence the power rates. To detect large cross-group differences, sample sizes of 100 or greater are needed to reach the power of 90% and above, which is the case for both the factor loadings and intercepts. However, detecting small differences requires much larger samples -- samples of 200 are large enough to detect small differences in the intercepts with powers greater than 80%, which is not large enough to detect small differences in the loadings, as the powers were around 65%. As shown, sample size of 500 produces power rates greater than 95% across all examined data conditions.

I compare the Type I error rates and power of locating non-invariant parameters between BSEM and ML-LRT approaches. Instead of performing LRTs on items as the standard ML-LRT does, I do it on specific tested parameters (i.e., intercepts and factor loadings). Specifically, a series of one-degree LRT are conducted between a baseline model (where equality constraints are only added on the RIs for identification purpose) and a constrained model (where equality constraint is added to each tested parameter). Results in Table 2 indicate that compared to ML-LRT, the BSEM approach is more conservative for sample sizes of 200 or less in detecting truly non-invariant parameters, demonstrating smaller Type I error rates but lower power. However, the two approaches show comparable Type I error rates and power when sample size is greater than 200.

3.3 Study III: The BSEM Approach in Evaluating the Consequences of Non-Invariance

Data Simulation

In study III, I aim to investigate the performance of using BSEM to evaluate the consequences of non-invariance. The data were generated based on a population two-group CFA model with five continuous indicators. For simplicity, I only considered the non-impact conditions where the reference group and focal group have the same population factor mean (e.g. 0.00), and variance (1.00). In the reference group, all factor loadings are simulated with a population value of 0.80, and all intercepts are simulated to be 0. In the focal group, the factor loadings and intercepts are set to be equivalent as the reference group (i.e. factor loadings equal 0.8 and intercepts equal 0) for the truly-invariant item (i.e. item 1). The rest four items are simulated to be non-invariant across groups. The four non-invariant items vary in terms of the source and magnitude of non-invariance. The population values for the non-invariant factor loadings and intercepts in the focal group are determined according to different simulation scenarios as described below. For both reference and focal groups, the population values for all residual variances are set to 0.36.

Three variables are manipulated in our data simulation: Sample size, source of non-invariance, and magnitude of non-invariance.

Sample size. The two groups are generated with equal number of observations. Sample sizes include 100, 200, 500, 1000, and 2000 per group.

Source of non-invariance. Non-invariance is simulated either on factor loadings or intercepts. All unique variances are simulated to be equivalent across groups.

Magnitude of non-invariance. Under conditions with small cross-group differences, factor loadings in the focal group decrease by 0.2, or intercepts increase by 0.3. Under the large difference conditions, factor loadings in the focal group decrease by 0.4, or intercepts increase by 0.6. The choices for the magnitudes are based on suggestions from previous literature (Kim & Yoon, 2011; Kim, Yoon & Lee, 2012; Meade & Lautenschlager, 2004). For each simulation scenario, 500 replications are generated and analyzed with Mplus 7.11(Muthén & Muthén, 1998-2012).

In order to evaluate the consequences of non-invariance, for each simulated scenarios, I select three different locations from the distributions of the latent traits or the observed test scores. To be representative to the entire statistical distribution, the three points are selected from the center (i.e. $z \approx 0.00$), tail (i.e. $z \approx +3.00$) and somewhere in middle (i.e. $z \approx +1.00$). Specifically, for EDOI and EDOT, I focus on conditions where the latent factor scores equal to 0.00, 1.00, or 3.00 (from a standardized normal distribution). EDLT are calculated for conditions where the (sum) test scores are 0.00, 5.00 and 15.00. In total, at the item level, sixty (5*2*2*3) different conditions are considered in the study. When the analyses are conducted at the test (scale) level, items with different sources and magnitudes of non-invariance are summed up to create the test score. As a result, the number of conditions at the test level is fifteen (5*3).

<u>Analysis</u>

Assuming all non-invariant parameters are correctly detected, partial invariant multiple-group CFA models are fitted; that is, equality constrains only added to truly invariant parameters, whereas all non-invariant parameters are freely estimated with non-informative priors. Each MCMC chain runs for a minimum of 50,000 and a maximum of 100,000 iterations, and only every 10th iteration is recorded to eliminate auto-correlations. For each condition, the EDOI, EDOT, and EDLT with corresponding 95% Bayesian Credible Interval (BCI) are obtained from the posterior distributions of the parameters created based on equations 8, 9, and 10.

The true values for EDOI, EDOT and EDLT can be obtained by plugging the population parameter values into the equations (8, 9, and 10). For each of the three measures (EDOI, EDOT, and EDLT), I compute the mean, standard deviation, and mean squared errors (MSE) across all 500 replications. In order to better compare the estimated values with the true population values, relative differences (RD) were also computed as an indicator of estimation bias (e.g.,Widaman, 1993; Hoogland & Boomsma, 1998; Pornprasertmanit, Lee, & Preaher, 2014). In the current study, RD was defined as following:

$$RD = \frac{\bar{\theta}_{est} - \theta_{true}}{\theta_{true}} \qquad (11)$$

Where $\bar{\theta}_{est}$ represents the mean of parameter estimates across all replications; θ_{true} indicates the true population value of the specific parameter. Therefore, RD evaluates the distance between the mean of estimates and the true population value relative to the population values, or the percentage of estimation bias with respect to the true values. In addition, for each condition, the 95% Coverage Rate (CR) is computed as the percent of replications where the 95% Bayesian Credible Interval (BCI) contains the true value.

Results

The results for EDOI and EDOT are summarized in table 5-8. In general, the Bayesian estimates for EDOI and EDOT are close to their true population values, except for cases where the sample size is small, and the latent traits are less likely to observed (i.e. from the tail of latent variable distributions). For example, when sample size is 100 and the latent trait are three standardized deviation above its mean, the RD for EDOT is underestimated by 32.60 percent (with MSE 1.65). Similar results are observed for EDLT where the sample size is small and the observed (total) score are away from its mean. As shown in table 9, when sample size is 100, and the observed (total) test score is 15, the mean square error (MSE) is 0.22. However, the 95% Bayesian Credible Intervals (BCI) well cover the true population values across all conditions considered in the study. As demonstrated in the tables, the 95% coverage rates are between 93%-97%, even for conditions where the estimates of effect size measures are biased.

Chapter 4 Empirical Study

In this chapter, I provide an empirical example to demonstrate the specific procedure of testing factorial invariance using the proposed Bayesian approach. First, an item with high likelihood to be truly-invariant is selected as RI. Second, based on the selected RI, non-invariant parameters are located. Finally, the consequences of non-invariance are investigated.

4.1 Measure and Data

I use the items from the Center for Epidemiologic Studies Depression Scale (CES-D, Radloff, 1977). Subjects are asked to indicate how often they have felt certain types of symptoms during the past week. Responses were made on a four-point Likert-type scale ranging from zero (Rarely or none of the time/Less than one day) to three (All of the time/5-7 days). The original version of the scale contains 20 items. In current study, a shortened (15-item), unidimensional version of the scale is used, as discussed in Edwards, Cheavens, Heiy and Cukrowicz (2010). The complete content of the measure are listed in Table 10.

Data was obtained from the China Family Panel Studies (CFPS), a nationally representative survey launched in 2010 by the Institute of Social Science Survey (ISSS) of Peking University (Xie & Hu, 2014). Only adults (16-65 years old) who responded on all 15 items were included in the analysis (N=26,841). The average age of participants was 40.98 years (SD=13.57 years). Males made up approximately 48.56% of the sample whereas females composed 51.44% of the sample. The means, standard deviations, as well as the correlations among responses of the 15 items are shown in Table 11. The Cronbach’s alpha using the full sample in the current study is 0.88.

4.2 Analyses and Results

4.21 Step 1: Selection of RI

The factorial invariance test is conducted on the shortened version of CES-D across genders. First, I select a RI using the method introduced in Analysis I. Specifically, Using Mplus 7.11, a multiple-group BSEM model is fitted by using commands “TYPE=MIXTURE” and “KNOWNCLASS” (Muthén & Asparouhov, 2012b for details). The difference measure (D) is defined as a difference between the same parameters across groups with the “NEW” option under “MODEL CONSTRAINT”. In order to properly set the scale for the latent variables in the multiple group models, informative priors with zero mean and small variance are introduced for all difference parameters (D) by using the “DIFF” option under “MODEL PRIOR”. To test the sensitivity to prior variances, I choose priors with four different values of variances, i.e., 0.0005, 0.001, 0.005 and 0.01, based on the results from the simulation experiment, which suggest that prior variances smaller than 0.01 are preferable in power. The minimum (i.e., 50, 000) and maximum (i.e., 100, 000) numbers of iterations for the MCMC chain are specified using “BITERATIONS”. The thinning of MCMC chain is assigned with the command “THIN”. A complete Mplus syntax for the step 1 analysis is available in Appendix B. From the Mplus output, information of the posterior distributions for $D_{\lambda j}$, and $D_{\tau j}$ can be obtained, from which the selection index (Δj) is then computed. Relevant results are summarized in Table 12.

According to the results, item 6 is identified as RI as using normal priors with variances 0.001 and 0.0005. When applying normally distributed priors with variances 0.005 and 0.01, item 3 is suggested to be RI. However, item 3 and item 6 produce two smallest values on Δj across all four different choices of priors. In the meanwhile, the selection indices for item 3 and 6 are noticeably smaller to those produced by other items. Therefore, both item 3 and 6 are eligible to be the RI. Besides produces small value on the selection index, by looking at its content representation, item 6 (i.e. I felt depressed) seems to be miniature of the measured construct (i.e. depression), which is considered as one of the guidelines for selecting RI (Karkee & Choi, 2005). Therefore, I select item 6 as the RI.

4.22 Step 2: Locating Non-invariant Parameters

Next, using item 6 as a RI, I further locate the non-invariant parameters based on the method discussed in Study II (Section 3.2). I fit a multiple-group BSEM model using the similar commands introduced above. However, instead of setting the scale of the latent variable by using informative priors, I use item 6 to identify the model and scale the latent variables by setting its factor loading and intercept equal across gender. Differences are examined for the remaining parameters by checking posterior distributions of the corresponding difference measures (D_{ij}), which is defined under “MODEL CONSTRAINT”. A complete Mplus syntax for the step 2 analysis is provided in Appendix C. The results are summarized in Table 13.

According to the results, most of the parameters are detected as non-invariant across genders. Item 3 is the only item found to be invariant in both factor loadings and intercepts, which is consistent with the finding from step 1 (i.e. item 3 is also suggested as RI in step 1). Other invariant parameters include factor loadings for item 1, 2, and 11, and intercepts for item 5.

4.23 Step 3: Evaluating the Consequences of Non-invariance

In step two, most of the parameters are concluded as non-invariant (i.e. 10/14 factor loadings and 12/14 intercepts), and therefore the legitimation of using the current scale for making cross gender comparison may be questionable. Nevertheless, providing that the number of observations included in the analysis is large, the detected non-invariance is likely to be caused by the trivial differences in parameter values, which makes no practical influences. Thus, I further evaluate the practical consequences of the non-invariant parameters.

First, I investigate the expected cross-group difference on the observed values given different latent factor scores. Specifically, I use the latent mean for the female group as the reference (i.e. zero) point; the latent space of interest contains all possible factor scores within ±4 standard deviations[6] around the reference point. Then 81 discrete points (i.e. scores) with equal interval (i.e. 0.1) are selected to approximate the continuous latent variable space. Based on the method introduced in Study III (Section 3.3), for each selected values of latent scores, the expected gender differences on observed scores, as well as the corresponding 95% BCI are obtained. The above mentioned analysis is conducted at both item- and scale- (i.e. total score) levels (i.e. EDOI & EDOT). For each of the 13 non-invariant items (except for items 3 and 6), the expected the differences on observed item scores (and the corresponding 95% BCIs) along the latent space is shown in Figures 1-13. Figure 14 demonstrate the expected gender differences on the total test scores (and the corresponding 95% BCIs) across the selected values on the latent factor. The relevant Mplus syntax for the above mentioned analysis is provided in Appendix D and Appendix E.

In addition, I explore the expected gender differences on the latent construct across a various values of the observed total scores. Since the scale has 15 items with 4 response categories (i.e. 0, 1.2 and 3), the possible observed total scores include 46 values (i.e. all integers from 0 to 45). Then the expected gender differences on the latent construct (with 95% BCI) for all 46 possible observed total scores are obtained using the method introduced in Study III (Section 3.3), and shown as Figure 15. The relevant Mplus syntax is provided in Appendix F.

At the item level, among the non-invariant items, item 1("I was bothered by things that usually don't bother me") and item 2 ("I did not feel like eating; my appetite was poor") produce very trivial influences in terms of observed differences conditional on the values of the latent variable. As shown in Figure 1 and Figure 2, since non-invariance only exist on intercepts, the lines indicate expected gender differences on observed scores across latent factors are parallel to the horizontal axis. Specifically, across the entire space of factor scores considered in the analysis, females tend to have larger observed score than males. However, the differences are very small. Given the same level of the latent factor, there are 95% confident that the gender differences for the observed scores are within the intervals [0.013, 0.048] (item 1), and [0.023, 0.055] (item2). For item 11 ("My sleep was restless"), females also always tend to produce about 0.1 points higher observed score (with 95% BCI [0.079, 0.117]).

For item 5, females tend to have higher observed scores when the level of the latent variable (i.e. depression) is below the group mean for females; whereas at high depression level (i.e. above the group mean for females), females produce lower observed scores in relative to males. At 95% confident level, the group differences on observed scores are less than 0.25 points across all levels of depression considered in the analysis. Similar pattern is observed for item 20; however, the threshold of depression level that suggests whether females have higher or lower observed scores (than males) is 0.50 (i.e. 0.5 stand deviations above the group mean for females).

For items 7, 9, 13, 14 and 19, females tend to give responses with smaller observed scores than males, unless when the level of depression is low[7]. Opposite patterns are observed for items 10 and 18, where females have higher observed scores than males except for individuals whose level of depression is low[8]. Roughly speaking, the gender differences on observed scores for the above mentioned items are less than 0.4 points across the entire latent variable space.

The non-invariance on item 17 ("I had crying spells") seems to make a great influence upon the conditional distributions of the observed response. Given the same level of depression, females tend to score higher on item 17 than males, except for individuals who have a low level of depression (i.e. 1.4 standard deviations below the group mean for female or less). In addition, the gender differences on observed scores can be relative large. For example, for individuals who suffer very high level of depression (3 standard deviation above the average depression level among females), females are expected to score 0.569 points higher on item 17 than males.

When calculating the total scores for all 15 items, as shown in Figure 14, females are expected to have higher total scores than males given the same level of depression. When depression is above the average level among females, the expected gender differences are small (i.e. about 0.10 points higher). However, for individuals whose depression symptom is below the average level among female population, females are expected to score about 0.6 points higher than males in total observed scores. It is also noted that the standard errors for the expected gender difference are larger as the latent scores (i.e. depression) becomes more extreme, and therefore yielding wider 95% credible interval.

Moreover, the expected gender differences on latent factor scores (i.e. depression) and corresponding 95% credible intervals across the space of observed total test scores are shown in Figure 15. Given an observed total test score, females are expected to have lower levels of the latent variable (i.e. depression) by about 0.01-0.02 points. The 95% credible intervals become wider as the observed scores are more deviant from its mean. Nevertheless, when considering the widest credible interval, the gender differences on latent scores are within ±0.1 points; the influence is still quite small compared to the scale of the latent variables in the analysis[9].

4.3 Summary of Major Findings

This study (in Chapter 4) works as a pedagogical example for demonstrating the proposed BSEM approach in studying factorial invariance using real data. In the meanwhile, empirically, the current study contributes to better understand the usage of CES-D scale across genders. The major findings are summarized as follows. 1.) Items 3 and 6 of the CES-D scale are (strong) invariant across genders. 2). Items 1, 2 and 11 are (weak) invariant across genders. (3). At the item level, the influence of the non-invariant items are generally small, except for item 17. Females tend to score higher on item 17 than males, unless the level of depression is low. (4). The non-invariance seems not to make a great influences at the total score level. For same level of depression, females tend to have higher observed total scores than males. However, the expected gender differences are less than 1 point (with 95% confident). In addition, given an observed total score, being female implies that lower level of depression than males; but the expected differences are small in relative to the metric of latent variable (depression).

Chapter 5 Discussion

5.1 Summary of Major Findings

In this dissertation, I propose using the BSEM approach for solving the three important issues in studying factorial invariance, including 1) how to select a proper RI, 2) how to locate non-invariant parameters given a RI has already been selected, and 3) how to evaluate the consequences of non-invariance.

The first step is to select a proper RI. To do so, informative priors with zero mean and small variance are imposed on the difference measure (D) for all loadings and intercepts. Items with the smallest value on the selection index (Δ_j) are then chosen as the RI. Once the RI is selected, the second step is to locate non-invariant parameters by examining the posterior distributions of all parameters except for those constrained to be exact equal for the RI item. If the 95% BCI of a parameter does not include zero, this parameter is then considered to be non-invariant. Finally, after the non-invariant parameters are detected, the non-invariance can be interpreted in terms of the expected differences in observed scores across levels of latent variable , or expected differences in latent traits conditioning on observed test scores (i.e. EDOI, EDOT, & EDLT). By using the information from the posterior distributions, the relevant confidence limits can be provided.

The BSEM approach for studying factorial invariance performs well with the simulated data. In the first step, it generally produces high power in detecting a truly invariant item. The power increases with decreasing proportion of non-invariance, increasing magnitude of non-invariance, and increasing sample size. In the second step, given that a truly invariant item is correctly chosen to be RI, the BSEM approach yields low Type I error rates in locating non-invariant parameters. Moreover, the power of this approach is high for sample sizes of 200 or greater, and is higher in locating non-invariance on intercepts than that on factor loadings. In the final step, for EDOI, EDOT and EDLT, the 95% Bayesian BCI shows well performance to cover the true population values across all simulated conditions considered in the study.

For the first two steps, I also compare the BSEM approach with two other approaches in addressing specification search problems. In identifying an invariant item to be RI, I compare BSEM with the constrained baseline approach. In locating non-invariant parameters given a correct RI, I compare BSEM with ML-LRT. The BSEM, the same as all other approaches, shows its own advantages and disadvantages. The BSEM approach greatly reduces workload by fitting one model in each step. It also performs sufficiently well across most of the investigated data conditions. The only downside I find is that small samples (200 or less) seem not preferable for BSEM in comparison with the other two. Specifically, compared to ML-LRT, the BSEM approach shows lower power in detecting truly non-invariant parameters, especially when sample size is 200 or less. This may limit its uses in studies with small samples, although sample size of 200 or more has long been suggested for typical SEM analysis in research. When sample size is more than 200, the BSEM method could demonstrate equivalent power as ML-IRT in locating non-invariant parameters.

In summary, as demonstrated in the empirical study, the proposed BSEM approach provides an alternative and complete procedure of conducting measurement invariance study, which could produce useful guidelines for applied researchers.

5.2 Prior Selection

An important feature of the BSEM approach is the use of informative priors. In practice, how to choose appropriate priors can be crucial in estimating any models using Bayesian methods (MacCallum, Edwards & Cai, 2012). Posterior distributions are obtained from modifying the likelihood using priors. Non-informative priors are often used when necessary population information is lacking. Since the prior carries little or no information about the parameter, the estimation is predominately determined by the data. In contrast, informative priors can reflect strong belief about parameters, thereby heavily influencing the posterior distributions. In Analysis I, I extend Muthén and Asparouhov (2012)'s idea of using informative priors of normal distribution with zero-mean and small-variance to replace the parameter specification of exact zero. These prior distributions do not directly reflect the researchers' prior knowledge and beliefs on the parameter of interests. Instead, the aim of utilizing the informative priors is to reach a goal that is not reachable using ML. That is, setting the difference parameters (D) close to zero so that the scale can be properly set. Meanwhile, by not strictly fixing those to zero, the difference parameters (D) can still be estimable and used as index for selecting RI.

As demonstrated in Little (2006), choosing different prior information could lead to different answers in Bayesian analysis. Consistent with previous simulation studies (e.g., Muthén & Asparouhov, 2013), study I (Section 3.1) also shows that choices of informative priors could influence the power of detecting invariant items. When variances of the priors are small enough for the difference measure, such as 0.01 and below, the power of correctly selecting an invariant item is consistently high and not sensitive to the choices of prior variances; however, when larger prior variances are used, such as 0.10, low power rates are observed under several conditions.

It has been suggested that if the variances of zero-mean priors are too small, the random draws in the posterior distributions are likely being pulled towards the zero prior mean; if the variances are too large, the random draws could become wild and produce undesirable estimates, and in some extreme cases, the models may become even unidentified (Muthén & Asparouhov, 2012). In the simulation studies, the prior with the largest variance of .10 is associated with less optimal results in terms of the power of detecting invariant items, compared with those obtained from priors with variances of .001, .01 and .05. Methodologists have recommended that sensitivity to prior variances should be examined by using priors with different variances in fitting BSEM models to real-world data (Muthén & Asparouhov, 2013). In the empirical analysis, we use informative priors with four different values of variances (i.e., 0.0005, 0.001, 0.005 and 0.01) in selecting an invariant item using the selection index. All four priors lead to the same item(s) being identified as invariant. Although it is not clear whether other unexamined prior variances would produce the same result, the proposed selection index seems to be not very sensitive to the choice of priors as long as proper prior variances are used. This may be due in part to the way how the proposed selection index works -- it summarizes the standardized difference across the factor loading and intercept for each item, and the item with the smallest value in this selection index is chosen to be the RI. In comparison with checking if zero lies in the 95% BCI of the posterior distribution for each individual cross-group difference, the selection index is less subject to the influence of the priors, because the RI is identified by comparing the magnitude of the selection index among all items.

It should be noted that the proposed BSEM method is different in several ways from the method by Muthén and Asparouhov (2013) introduced as web notes, although they share certain similarities in terms of using informative priors. The goal of Muthén and Asparouhov (2013) was to compare factor means and variances across many groups or time points. For this purpose, they proposed a two-step procedure implemented by BSEM approach – step 1 to identify non-invariant parameters and step 2 to free those parameters, thereby estimating factor means and variances with approximate measurement invariance. Later they extended this idea and proposed the so-called alignment method to estimate group-specific factor means and variances with approximate measurement invariance (Asparouhov & Muthén, 2014). However, the goal in the first two studies in this dissertation is to solve two common issues in measurement invariance tests – how to select appropriate reference indicators and then how to locate non-invariant parameters, given the metric of latent constructs and other parameters is appropriately set by the reference indicators chosen in the first step. The first step in Muthén and Asparouhov (2013)'s method is to identify *non-invariant parameters* without using any reference indicator, but the step 1 in current study is to identify an *item* with highest likelihood to be *invariant* and then use such item as a reference indicator in subsequent analysis. Both methods use zero-mean and small-variance priors; however, the selection criteria are constructed differently. The proposed method combines the standardized cross-group differences in both factor loadings and intercepts for each item, but their method focuses on raw difference between each individual parameter and the group-mean of that parameter. These two methods may perform similarly in terms of identifying non-invariant parameters (what step 2 is for in the current study) under certain circumstances. However, systematic comparisons between these methods are out of scope of the current investigation.

5.3 Selection Index Using Factor Loading Only

As discussed earlier, there are different levels of factorial invariance. In practice, strong factorial invariance is required in many situations, such as comparing latent/observed means (Steinmetz, 2013) and fitting latent growth models (Ferrer, Balluerka, & Widaman, 2008). Therefore, I focus on selecting RI with both equal factor loadings and equal intercepts, and consider items which are invariant in both factor loadings and intercepts as truly invariant items. However, if one is only interested in testing for weak invariance, items with equal factor loadings only can be considered as truly invariant. Accordingly, a simplified selection index, which includes information on factor loadings only, can be used. The modified section index can be expressed as:

$$\hat{\Delta}_j{}' = \frac{|\hat{\mathrm{D}}_{\lambda\mathrm{j}}|}{sd_{\lambda j}} \qquad (12)$$

An additional simulation study is conducted to investigate the performance of the modified BSEM method on selecting an RI with invariant factor loading only. The data simulation procedure and population parameters settings are the same as Study I (Section 3.1). Specifically, I only consider the simulation conditions where non-invariance only exists on factor loadings. The RI is selected using the BSEM method with the modified selection index. Two different informative priors (i.e. Prior1~N (0, 0.001) and Prior2~N (0, 0.01)) are used to properly scale the latent variables in the multiple group model. The power rates of the modified BSEM method across simulation conditions were reported in Table 14. The results show that the BSEM method still works well with the modified selection index. The power rates for correctly selecting an RI are larger than 93% as the sample size exceeds 200. Therefore, the proposed BSEM methods can be easily transformed to fit situations where researchers need to select RI with equal factor loadings only.

5.4 Revisiting the Comparison between BSEM and the Constrained Baseline Approaches on Selecting RI

In study I (Section 3.1), the proposed BSEM method is found to be superior to the constrained baseline approach under several simulation conditions. Under conditions where the amount or percent of non-invariance was large, as sample size increase, the constrained baseline approach tended to reject the truly-invariant items during the screening phase. For example, when 40% of intercepts were contaminated with large amount of non-invariance; as sample size reached 500, the constrained baseline approach produced a power rate of zero by rejecting every item as invariant.

However, when the sample size is small (i.e. less than 200), the constrained baseline model seemed to gain higher power compared to the BSEM approach. As discussed hereinbefore, the power rates for the constrained baseline approach reported in Study I (Section 3.1) may be overestimated caused by the simulation design. That is, in the simulation study I, for the non-invariant item(s), their factor loadings decreased in the focal group. As a result, when adding the equality constraints, the truly invariant items happen to have larger factor loadings, compared to those for non-invariant items in our simulated data. In this case, the constrained baseline approach, characterized by selecting an item with non-significant cross-group difference but the largest loading to be RI, would tend to have relatively high power to choose a truly invariant item to be RI.

In order to investigate the possible overestimation of power rates for the constrained baseline approach caused by the simulation design, I conduct an additional simulation to further compare the BSEM and constrained baseline approaches on selecting RI. For simplicity, I only consider cases where small amount of non-invariance existed on factor loadings. In addition, the factor loadings for non-invariant items increased in the focal group. The population values of factor loadings for invariant items are set to be 0.6 in both reference and focal groups. However, for the non-invariant items, the factor loadings for the reference group are set to be 0.8; for the focal group, the population values for factor loadings are still 0.6. The set-up for other parameters is the same as discussed in Section 3.1. The power rates for the proposed BSEM method across simulation conditions are reported in Table 15, and compared to the power of the constrained baseline approach.

In the additional simulation, the BSEM method showed better performance than the constrained baseline approach across all conditions, including conditions where the sample size is small. For example, under the condition that 20% of the factor loadings are with small amount of non-invariance, when sample size is 50, the power rate for correctly selecting RI is greater than 90% by using the BSEM approach. However, the constrained baseline approach only yields 56% probability to select a correct RI.

5.5 Extensions

A few other possible extensions could be made to widen the uses of the BSEM approach through borrowing lessons from the IRT literature. One lesson is that among items with similar non-significant cross-group differences, choosing items with the strongest relationship with the factor would produce greater power in subsequent identification of non-invariant parameters (Meade & Wright, 2012). Another lesson is that a large number of studies have suggested that using more than one RI results in greater power in subsequent invariance tests (Meade & Wright, 2012). In the SEM literature, however, much less is known about whether the above mentioned conclusions hold for multiple-group CFA analysis. Therefore, I conduct additional simulations to verify the effects of number of RI and the magnitude of factor loadings on the power to detect non-invariance within the SEM framework. Two additional simulation conditions were considered.

1.) Number of RI in use (One RI vs. Three RIs)

2.) Magnitude of the selected RI's factor loading (0.8 vs. 0.4)

The two additional conditions were then fully nested to the existing simulation scenarios discussed on Section 3.2, resulting in 256 (2*2*64) conditions. The type I error rates and power rates for both BSEM and likelihood ratio tests are calculated and reported in tables 16-19. As showed in Table 16-19, for both BSEM method and likelihood ratio test under MLE, higher power rates to detecting non-invariant parameters are observed when more RI(s) are included, and(or) item with larger factor loading is used as RI. For example, for the BSEM method, the power rate for correctly detecting small amount of non-invariance on factor loadings (2/10 factor loadings are non-invariance) with sample size 200 is 0.12 using only one RI with population factor loadings 0.4. However, the power increases to 0.56 when using one RI with population factor loading 0.8. Using three RIs with population factor loadings 0.4 also improves the power to 0.50. The largest power rate (i.e. 0.80) is observed when using multiple (i.e. three) RIs with larger factor loadings (i.e.0.8). In addition, for the BSEM method, when sample size was small, more RI and larger factor loadings of RI leaded to higher admissible solution rates. Therefore, in order to achieve higher power rate, it is more ideal to select more than one RI and RI with larger factor loadings. I further explore the possibilities to extend the BSEM method in order to incorporate the above conclusions.

5.5.1 Selecting More than One RI(s)

According to the BSEM approach, small selection index indicating high likelihood for an item to be truly invariant. Therefore, a natural idea for selecting multiple RIs is to select a subset of items which produce smallest values on the selection index. In the current section, I conduct simulation study to investigate the performance of the modified BSEM method on selecting more than one (i.e. three) RIs. For simplicity, I focus on selecting RIs with non-invariant factor loadings only. The set-up for parameter values and simulations conditions are the same as Section 5.4. For the BSEM method, three items with smallest selection indices were selected as RIs. The power rates of the modified BSEM method were reported in Table 20, and compared to the constrained baseline approach.

As showed in table 20, the power rates for the modified BSEM method to select three RIs are superior to the traditional constrained baseline approach across all simulated conditions. The BSEM method performs well especially under conditions where the percentage of non-invariance is low. For example, when percentage of non-invariance is low, the BSEM method has more than 90% probability to selecting all three RIs correctly if the sample size reaches 200.

5.5.2 Selecting RI with Larger Communalities

In this section, I investigate a possible way to increase the chance for the BSEM method to select a truly invariant item with larger factor loadings as RI. According to property 2, the non-invariant item(s) are expected to be associated with noticeably larger selection indices ($\hat{\Delta}_j$), compared to those produced by the truly invariant items. The differences of the magnitude of selection indices ($\hat{\Delta}_j$) among truly invariant items, however, are anticipated to be much smaller, and only subject to the random errors. Given two truly invariant items, it is possible that one with smaller factor loadings produced the smallest selection index; the other item with higher factor loadings, even has a fairly close (but slightly larger) selection index, is not selected as RI. In order to distinguish the “large” difference on selection index caused by non-invariance and the variabilities due to random errors, an empirical tolerance level (TOL) is proposed and can be expressed as

$$TOL = \frac{Max(\Delta_j) - Min(\Delta_j)}{c} \qquad (13)$$

The tolerance index takes the difference between the largest and the smallest selection index, and divides it by a constant *c*. If several items produced selection indices ($\hat{\Delta}_j$) which are similar to the smallest one, such that the differences are smaller than the pre-determined tolerance level, or

$$\hat{\Delta}_j < Min(\hat{\Delta}_j) + TOL, \qquad (14)$$

item with higher factor loadings is suggested as RI. The choice of the tolerance level (thus value of *c*) could be subjective; the tolerance level is expected to be large enough to count for the variabilities of selection index among truly invariant items. In the meanwhile, it should not be too inflated; otherwise the non-invariant item(s) would not be detectable.

I further investigate the performance of using TOL to select RI with larger communalities using a small scaled simulation. For simplicity, I only consider the cases of selecting RI with invariant factor loadings. In order to evaluate the performance of the modified BSEM method on selecting RI with the larger factor loadings, I further manipulate the magnitude of factor loadings for the truly invariant items. Specifically, under 5-item conditions, when the percent of non-invariance was large, we set the factor loadings for the three truly invariant items to be 0.4, 0.6 and 0.8 respectively; for cases with small percent of non-invariance, half of the invariant items had population factor loadings 0.8, whereas the other half had factor loadings of 0.4. When the total number of item was 10, half of the factor loadings for truly invariant items were set to be 0.8, whereas population loadings for the other half were set as 0.4. The set-up for other parameters and simulation conditions can be referred to the Section 3.1. In this pilot analysis, I used c=2 based on a rough observation of the empirical distributions of the selection index between non-invariant and invariant items. Among the items that produced selection index satisfied:

$$\hat{\Delta}_j < Min(\hat{\Delta}_j) + \frac{Max(\Delta_j) - Min(\Delta_j)}{2}, \tag{15}$$

I select the item with largest factor loading as RI. The power rates for the modified BSEM method on selecting RI with different factor loadings were reported in Table 21.

As shown in Table 21, in most of the cases, using selection index with tolerance level increase successfully increase the chance for the BSEM method to select an invariant item with relative larger factor loadings as RI, while still keep a low level of Type I error rate (i.e. chance of incorrectly choosing RI). For example, when 4 out of 10 factor loadings are infected with large magnitude of non-invariance with sample size 200, the BSEM method using tolerance level yields 90% probability to correctly chose invariant items with higher factor loading (i.e. 0.8) as RI. The probability of correctly choosing an RI which is truly invariant but with smaller factor loading (i.e. 0.4) is much lower (i.e. less than 3%); so is the probability of incorrectly choosing a non-invariant item as RI (less than 7%). Nevertheless, when 2 out of 5 factor loadings are infected with small magnitude of non-invariance, using selection index with tolerance level could increase the chance of making type I error rate, especially when sample size is small; for example, when sample size is 50, the type I error rate is inflated to more than 30%.

In summary, in the Section 5.5, I heuristically explore several possible extensions to the proposed selection index. These extensions can facilitate researchers to select RI which is not only truly-invariant, but also possible to generate higher power in the subsequent analysis for locating non-invariant parameters. Using small scaled simulations, these extensions could show promising performance to select more than one RIs and RI with larger factor loadings. Future studies are expected to further investigate and develop upon these extensions in a more systematic manner and thus better resolve the specification search issues in studying factorial invariance.

5.6 Supporting Null Hypothesis

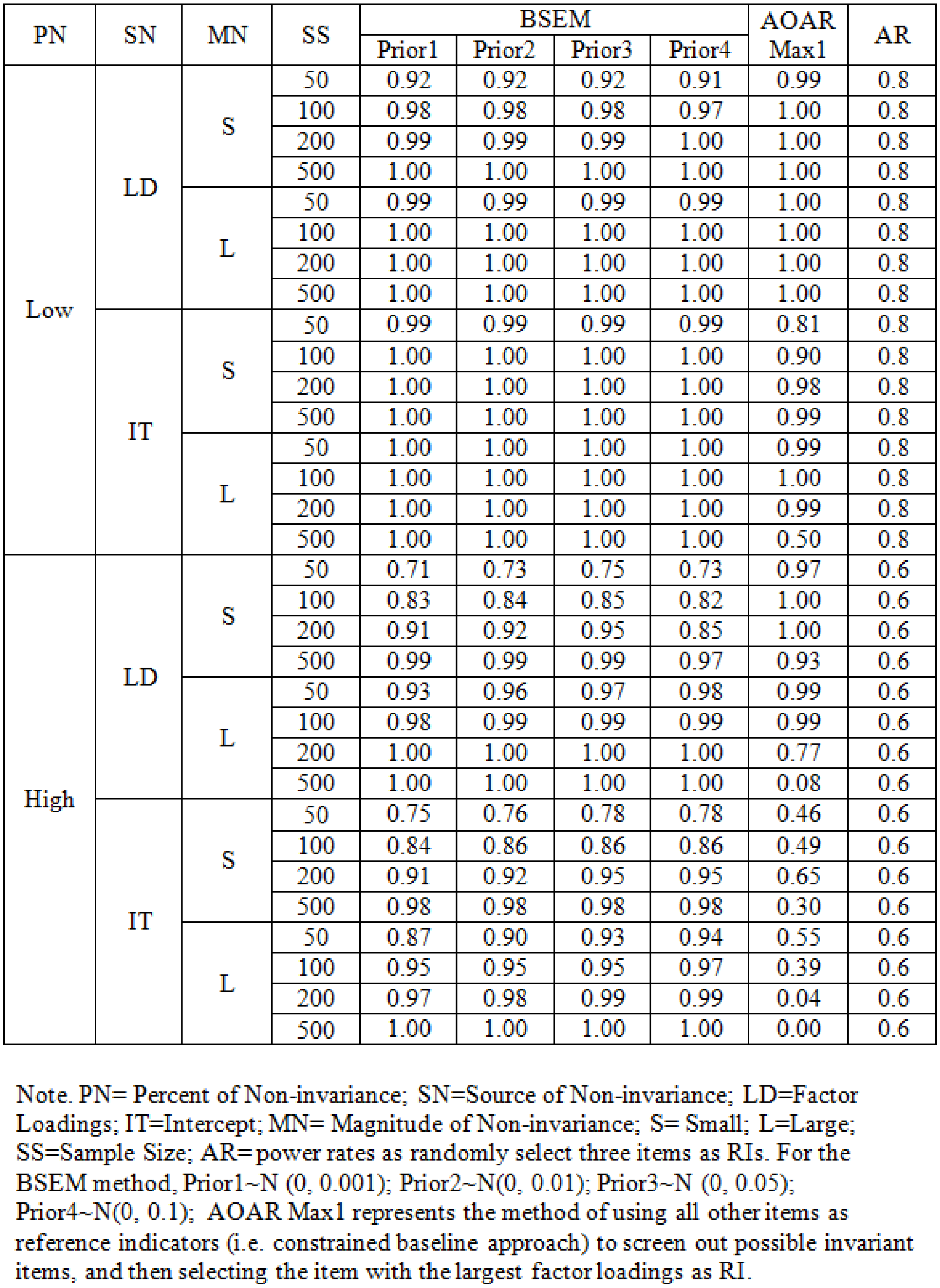
By obtaining the complete posterior distribution, more can be learned from the parameter estimates using the Bayesian methods. Therefore, the Bayesian approach could provide more useful information in studying measurement invariance, comparing to the traditional ML methods. Specifically, if the ML-based LRT test turns out to be non-significant, researchers tend to conclude no cross-group differences on the tested parameters (i.e. factorial invariance holds across groups). Despite often reported by applied researchers (e.g. Liu, Borg, & Spector, 2004; Shevlin & Adamson, 2005), the above statement is subject to one of the most common misinterpretations in the framework of null hypothesis significance test. That is, failing to reject provides no basis about accepting the null hypothesis (Cohen, 1994).

The Bayesian methods, however, provide possible tools for researchers to accept a null value, and thereby offer direct evidence for supporting factorial invariance. For example, Verhagen, Levy, Millsap and Fox (2015) proposed tests based on Bayes factors to evaluate the evidence in favor of the null hypothesis of invariance in IRT models. Another possible way is to establish a region of practical equivalence (ROPE) around zero for the difference parameter (D). For example, suppose researchers decide that any absolute differences less than 0.05 in factor loadings could make no practical significant (i.e. can be practically treated as invariant). Therefore, the ROPE for the difference parameter (D) on factor loadings is [-0.05, 0.05]. If the 95% BCI falls completely inside the ROPE, researchers would more confidently conclude that factorial invariance holds for factor loadings, because the 95 percent of the most credible values for the difference in factor loadings are practically zero. These advantages of using the Bayesian framework could lead possible extensions for using BSEM to study factorial invariance, which deserve more attention for future investigations.

5.7 Limitation and Other Future Directions

Some limitations need to be noted about this study. In the simulation analyses, I assume the direction of non-invariance is uniform, that is, the non-invariant loadings or intercepts always take higher values for one group than the other. Since all non-invariant parameters have a uniform direction across groups, using all parameters (as reference) to scale the latent variable could produce poor approximation to the appropriate metric that is otherwise set by using truly invariant parameters only. Things could be even worse if the non-invariance is also proportional in magnitude (Yoon & Millsap, 2007). Under this condition, the non-invariant loadings would share a common (biased) metric, and impacts of the non-invariant loadings are expected to be superposed in the adjusted metric of latent constructs. The more favorable type of non-invariance would be mixed in direction of non-invariance, in which one group has greater values on some of the parameters but lower values on the others than other groups. Since the non-invariant parameters differ across groups with opposite directions, adverse impact of the non-invariance on the adjusted metric of latent constructs could be counteracted. Second, I do not investigate the performance of BSEM in the case where both intercepts and factor loadings are different for items at the same time. I anticipate that the selection index would be much more effective in this case because the overall cross-group difference should be greater than that for the case with either one to be different. Cases like this are worthy of systematic examination in the future. Third, in current study, I first focus on selecting RIs and then assuming the RIs are correctly selected, I investigate the power of locating non-invariant parameters. Apparently, these two steps are separate in our design, that is, the results of Analysis II are not conditional on Analysis I. In practice, if an error is made in the first step (i.e. select a wrong RI), the performance of the BSEM method to locate non-invariant parameters requires further investigations. Finally, in the simulation analyses, I consider only the situation where the majority of parameters were invariant, as the case for most well-developed instruments. I show that the proposed BSEM approach still performed reasonably well under conditions of uniform non-invariance. However, it is not clear how this approach would behave in selecting invariant and detecting non-invariant parameters if the majority of parameters are non-invariant, uniform in direction, and proportional in magnitude. More research would be needed to investigate such extreme cases.

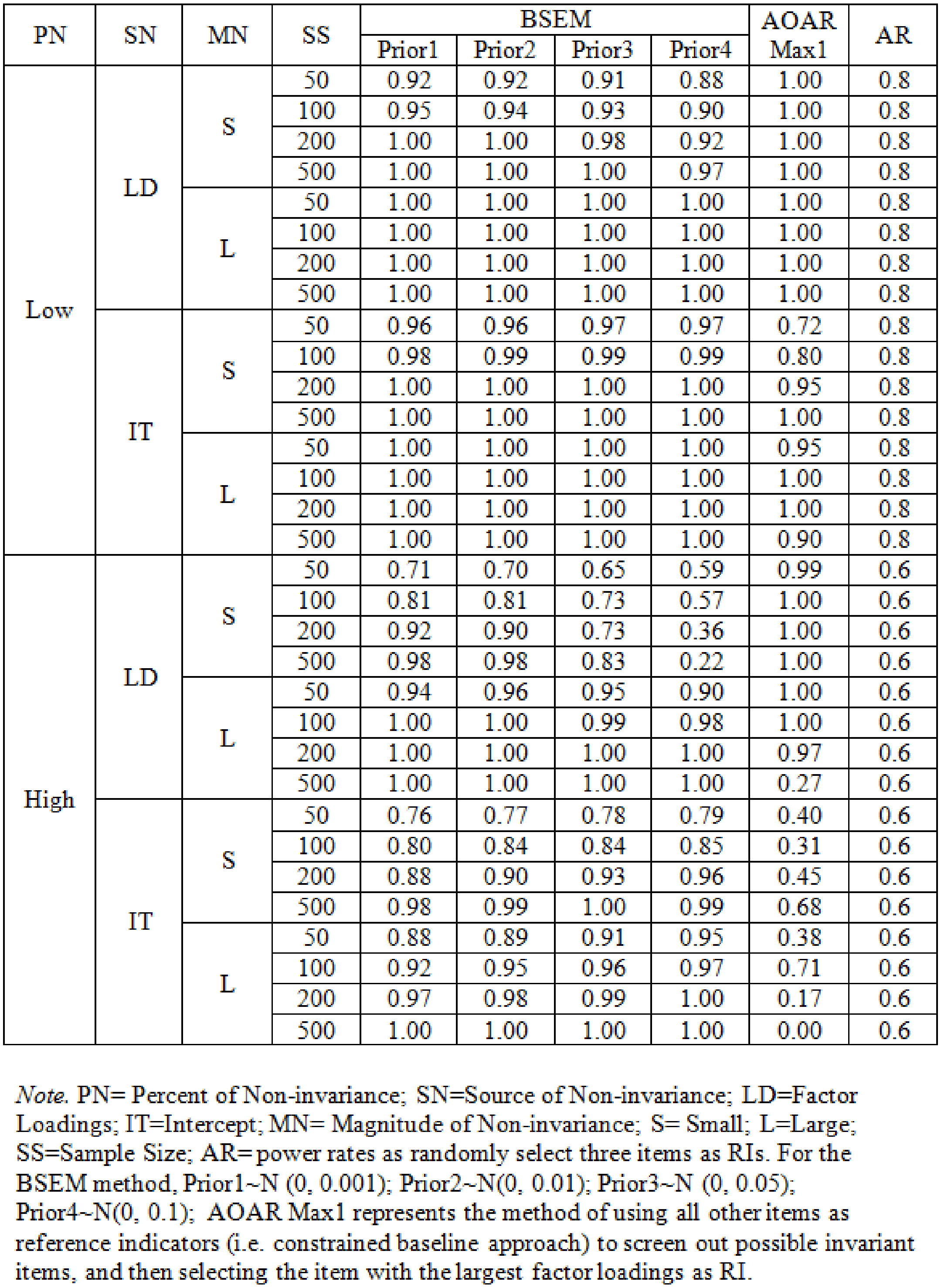
Table 1: Power Rates of Selecting a Reference Indicator (5 Items)



PN	SN	MN	SS	BSEM				AOAR	AR
				Prior1	Prior2	Prior3	Prior4	Max1	
Low	LD	S	50	0.92	0.92	0.92	0.91	0.99	0.8
			100	0.98	0.98	0.98	0.97	1.00	0.8
			200	0.99	0.99	0.99	1.00	1.00	0.8
			500	1.00	1.00	1.00	1.00	1.00	0.8
		L	50	0.99	0.99	0.99	0.99	1.00	0.8
			100	1.00	1.00	1.00	1.00	1.00	0.8
			200	1.00	1.00	1.00	1.00	1.00	0.8
			500	1.00	1.00	1.00	1.00	1.00	0.8
	IT	S	50	0.99	0.99	0.99	0.99	0.81	0.8
			100	1.00	1.00	1.00	1.00	0.90	0.8
			200	1.00	1.00	1.00	1.00	0.98	0.8
			500	1.00	1.00	1.00	1.00	0.99	0.8
		L	50	1.00	1.00	1.00	1.00	0.99	0.8
			100	1.00	1.00	1.00	1.00	1.00	0.8
			200	1.00	1.00	1.00	1.00	0.99	0.8
			500	1.00	1.00	1.00	1.00	0.50	0.8
High	LD	S	50	0.71	0.73	0.75	0.73	0.97	0.6
			100	0.83	0.84	0.85	0.82	1.00	0.6
			200	0.91	0.92	0.95	0.85	1.00	0.6
			500	0.99	0.99	0.99	0.97	0.93	0.6
		L	50	0.93	0.96	0.97	0.98	0.99	0.6
			100	0.98	0.99	0.99	0.99	0.99	0.6
			200	1.00	1.00	1.00	1.00	0.77	0.6
			500	1.00	1.00	1.00	1.00	0.08	0.6
	IT	S	50	0.75	0.76	0.78	0.78	0.46	0.6
			100	0.84	0.86	0.86	0.86	0.49	0.6
			200	0.91	0.92	0.95	0.95	0.65	0.6
			500	0.98	0.98	0.98	0.98	0.30	0.6
		L	50	0.87	0.90	0.93	0.94	0.55	0.6
			100	0.95	0.95	0.95	0.97	0.39	0.6
			200	0.97	0.98	0.99	0.99	0.04	0.6
			500	1.00	1.00	1.00	1.00	0.00	0.6

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR=power rates as randomly select three items as RIs. For the BSEM method, Prior1~N (0, 0.001); Prior2~N(0, 0.01); Prior3~N (0, 0.05); Prior4~N(0, 0.1); AOAR Max1 represents the method of using all other items as reference indicators (i.e. constrained baseline approach) to screen out possible invariant items, and then selecting the item with the largest factor loadings as RI.

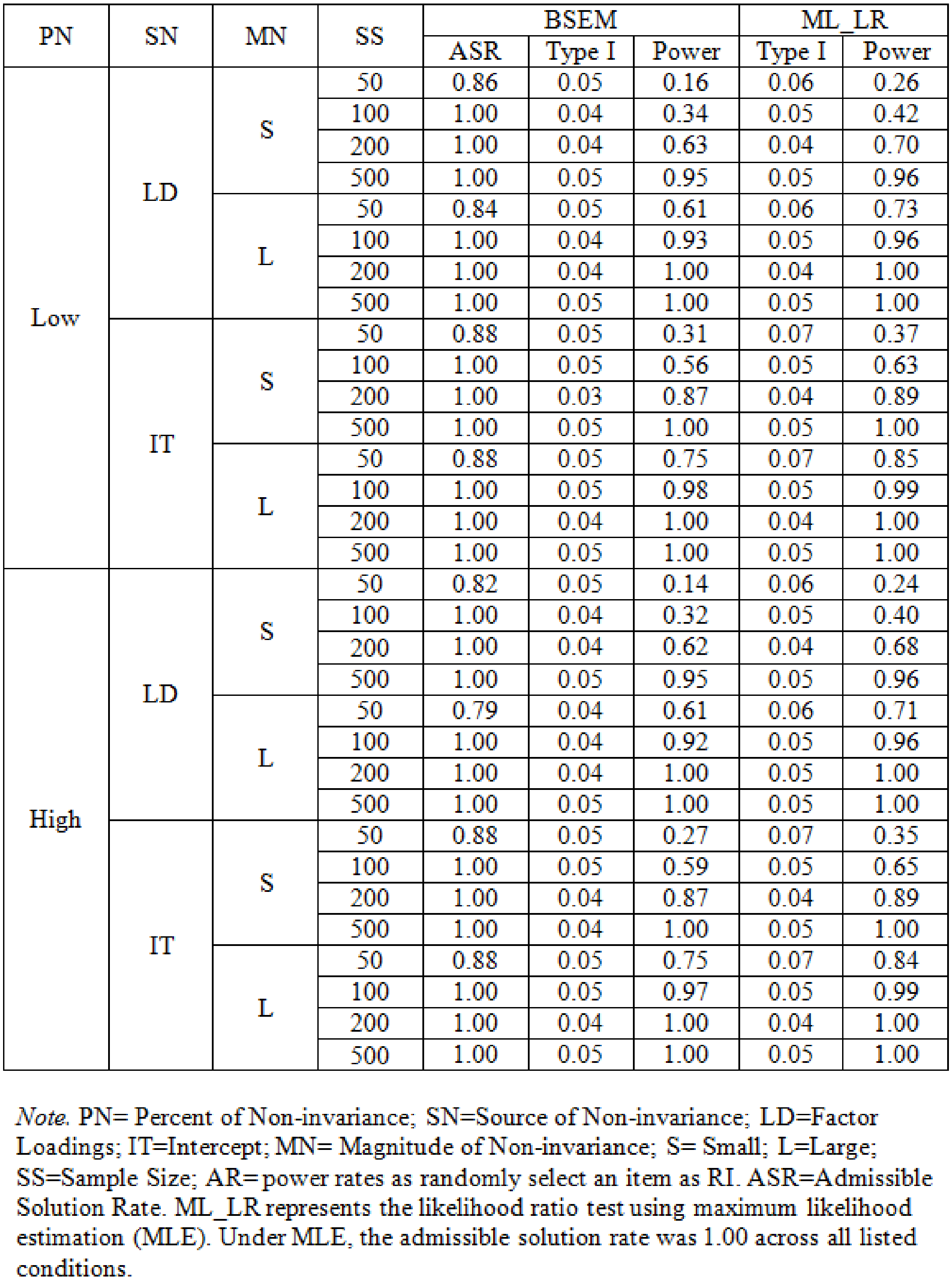
Table 2: Power Rates of Selecting a Reference Indicator (10 Items)



PN	SN	MN	SS	BSEM				AOAR	AR
				Prior1	Prior2	Prior3	Prior4	Max1	
Low	LD	S	50	0.92	0.92	0.91	0.88	1.00	0.8
			100	0.95	0.94	0.93	0.90	1.00	0.8
			200	1.00	1.00	0.98	0.92	1.00	0.8
			500	1.00	1.00	1.00	0.97	1.00	0.8
		L	50	1.00	1.00	1.00	1.00	1.00	0.8
			100	1.00	1.00	1.00	1.00	1.00	0.8
			200	1.00	1.00	1.00	1.00	1.00	0.8
			500	1.00	1.00	1.00	1.00	1.00	0.8
	IT	S	50	0.96	0.96	0.97	0.97	0.72	0.8
			100	0.98	0.99	0.99	0.99	0.80	0.8
			200	1.00	1.00	1.00	1.00	0.95	0.8
			500	1.00	1.00	1.00	1.00	1.00	0.8
		L	50	1.00	1.00	1.00	1.00	0.95	0.8
			100	1.00	1.00	1.00	1.00	1.00	0.8
			200	1.00	1.00	1.00	1.00	1.00	0.8
			500	1.00	1.00	1.00	1.00	0.90	0.8
High	LD	S	50	0.71	0.70	0.65	0.59	0.99	0.6
			100	0.81	0.81	0.73	0.57	1.00	0.6
			200	0.92	0.90	0.73	0.36	1.00	0.6
			500	0.98	0.98	0.83	0.22	1.00	0.6
		L	50	0.94	0.96	0.95	0.90	1.00	0.6
			100	1.00	1.00	0.99	0.98	1.00	0.6
			200	1.00	1.00	1.00	1.00	0.97	0.6
			500	1.00	1.00	1.00	1.00	0.27	0.6
	IT	S	50	0.76	0.77	0.78	0.79	0.40	0.6
			100	0.80	0.84	0.84	0.85	0.31	0.6
			200	0.88	0.90	0.93	0.96	0.45	0.6
			500	0.98	0.99	1.00	0.99	0.68	0.6
		L	50	0.88	0.89	0.91	0.95	0.38	0.6
			100	0.92	0.95	0.96	0.97	0.71	0.6
			200	0.97	0.98	0.99	1.00	0.17	0.6
			500	1.00	1.00	1.00	1.00	0.00	0.6

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR=power rates as randomly select three items as RIs. For the BSEM method, Prior1~N (0, 0.001); Prior2~N(0, 0.01); Prior3~N (0, 0.05); Prior4~N(0, 0.1); AOAR Max1 represents the method of using all other items as reference indicators (i.e. constrained baseline approach) to screen out possible invariant items, and then selecting the item with the largest factor loadings as RI.

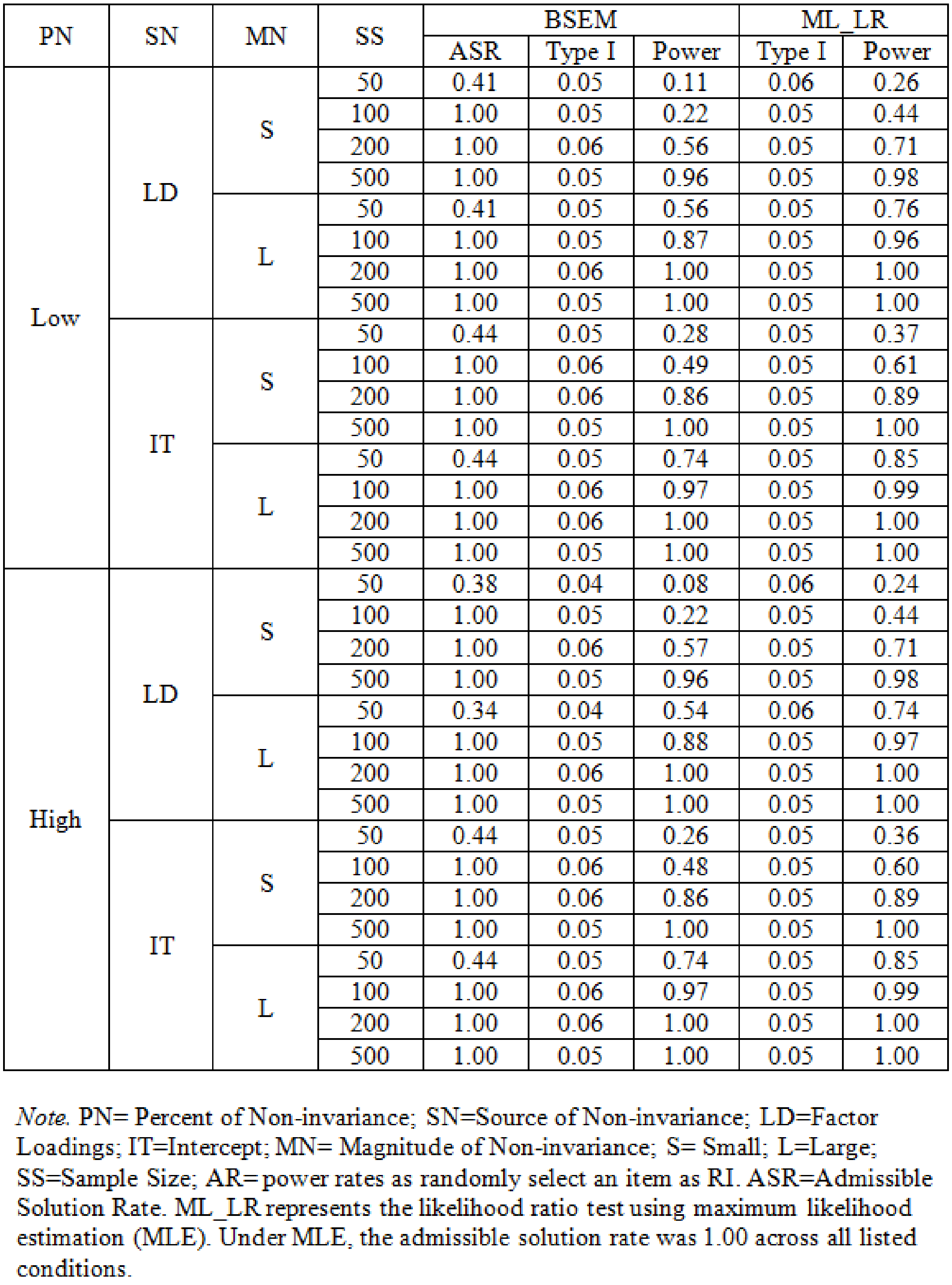
Table 3: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (5 Item)



PN	SN	MN	SS	BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power
Low	LD	S	50	0.86	0.05	0.16	0.06	0.26
			100	1.00	0.04	0.34	0.05	0.42
			200	1.00	0.04	0.63	0.04	0.70
			500	1.00	0.05	0.95	0.05	0.96
		L	50	0.84	0.05	0.61	0.06	0.73
			100	1.00	0.04	0.93	0.05	0.96
			200	1.00	0.04	1.00	0.04	1.00
			500	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.88	0.05	0.31	0.07	0.37
			100	1.00	0.05	0.56	0.05	0.63
			200	1.00	0.03	0.87	0.04	0.89
			500	1.00	0.05	1.00	0.05	1.00
		L	50	0.88	0.05	0.75	0.07	0.85
			100	1.00	0.05	0.98	0.05	0.99
			200	1.00	0.04	1.00	0.04	1.00
			500	1.00	0.05	1.00	0.05	1.00
High	LD	S	50	0.82	0.05	0.14	0.06	0.24
			100	1.00	0.04	0.32	0.05	0.40
			200	1.00	0.04	0.62	0.04	0.68
			500	1.00	0.05	0.95	0.05	0.96
		L	50	0.79	0.04	0.61	0.06	0.71
			100	1.00	0.04	0.92	0.05	0.96
			200	1.00	0.04	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.88	0.05	0.27	0.07	0.35
			100	1.00	0.05	0.59	0.05	0.65
			200	1.00	0.04	0.87	0.04	0.89
			500	1.00	0.04	1.00	0.05	1.00
		L	50	0.88	0.05	0.75	0.07	0.84
			100	1.00	0.05	0.97	0.05	0.99
			200	1.00	0.04	1.00	0.04	1.00
			500	1.00	0.05	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR=power rates as randomly select an item as RI. ASR=Admissible Solution Rate. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed conditions.

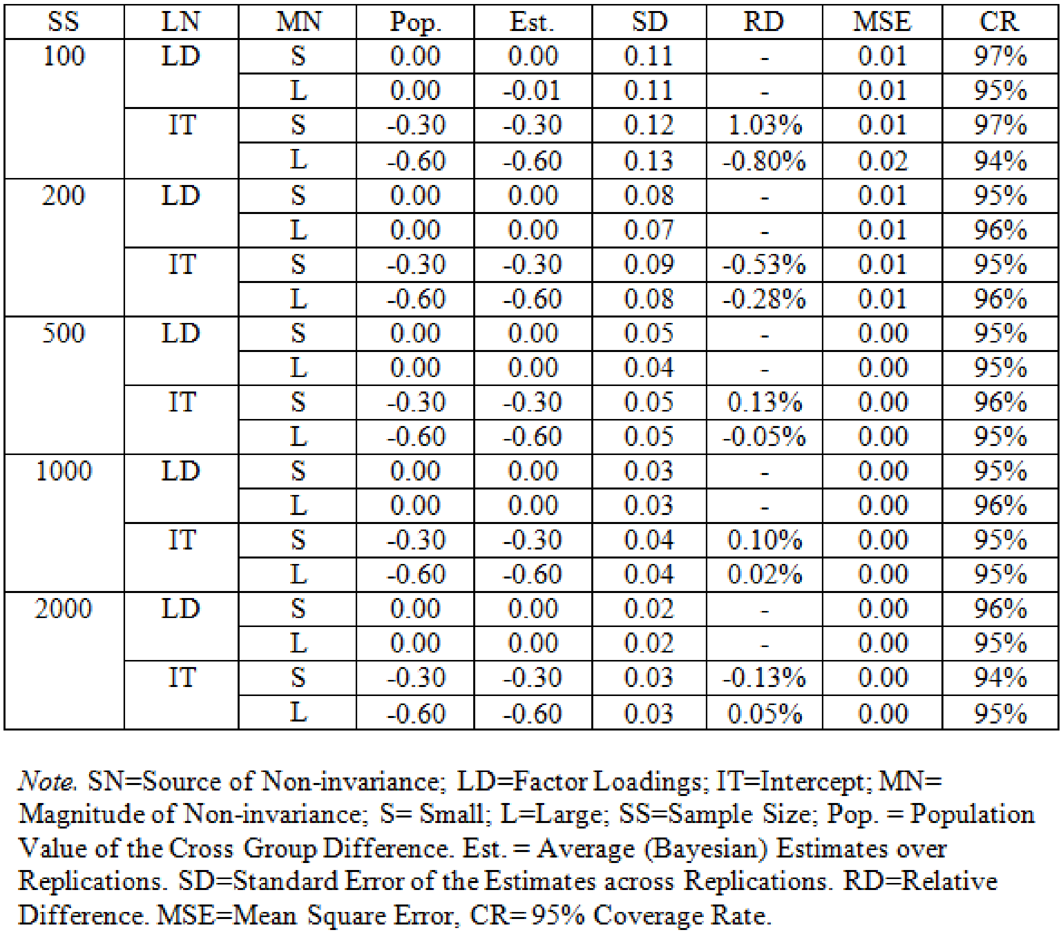
Table 4: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (10 Item)



PN	SN	MN	SS	BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power
Low	LD	S	50	0.41	0.05	0.11	0.06	0.26
			100	1.00	0.05	0.22	0.05	0.44
			200	1.00	0.06	0.56	0.05	0.71
			500	1.00	0.05	0.96	0.05	0.98
		L	50	0.41	0.05	0.56	0.05	0.76
			100	1.00	0.05	0.87	0.05	0.96
			200	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.44	0.05	0.28	0.05	0.37
			100	1.00	0.06	0.49	0.05	0.61
			200	1.00	0.06	0.86	0.05	0.89
			500	1.00	0.05	1.00	0.05	1.00
		L	50	0.44	0.05	0.74	0.05	0.85
			100	1.00	0.06	0.97	0.05	0.99
			200	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00
High	LD	S	50	0.38	0.04	0.08	0.06	0.24
			100	1.00	0.05	0.22	0.05	0.44
			200	1.00	0.06	0.57	0.05	0.71
			500	1.00	0.05	0.96	0.05	0.98
		L	50	0.34	0.04	0.54	0.06	0.74
			100	1.00	0.05	0.88	0.05	0.97
			200	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.44	0.05	0.26	0.05	0.36
			100	1.00	0.06	0.48	0.05	0.60
			200	1.00	0.06	0.86	0.05	0.89
			500	1.00	0.05	1.00	0.05	1.00
		L	50	0.44	0.05	0.74	0.05	0.85
			100	1.00	0.06	0.97	0.05	0.99
			200	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR=power rates as randomly select an item as RI. ASR=Admissible Solution Rate. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed conditions.

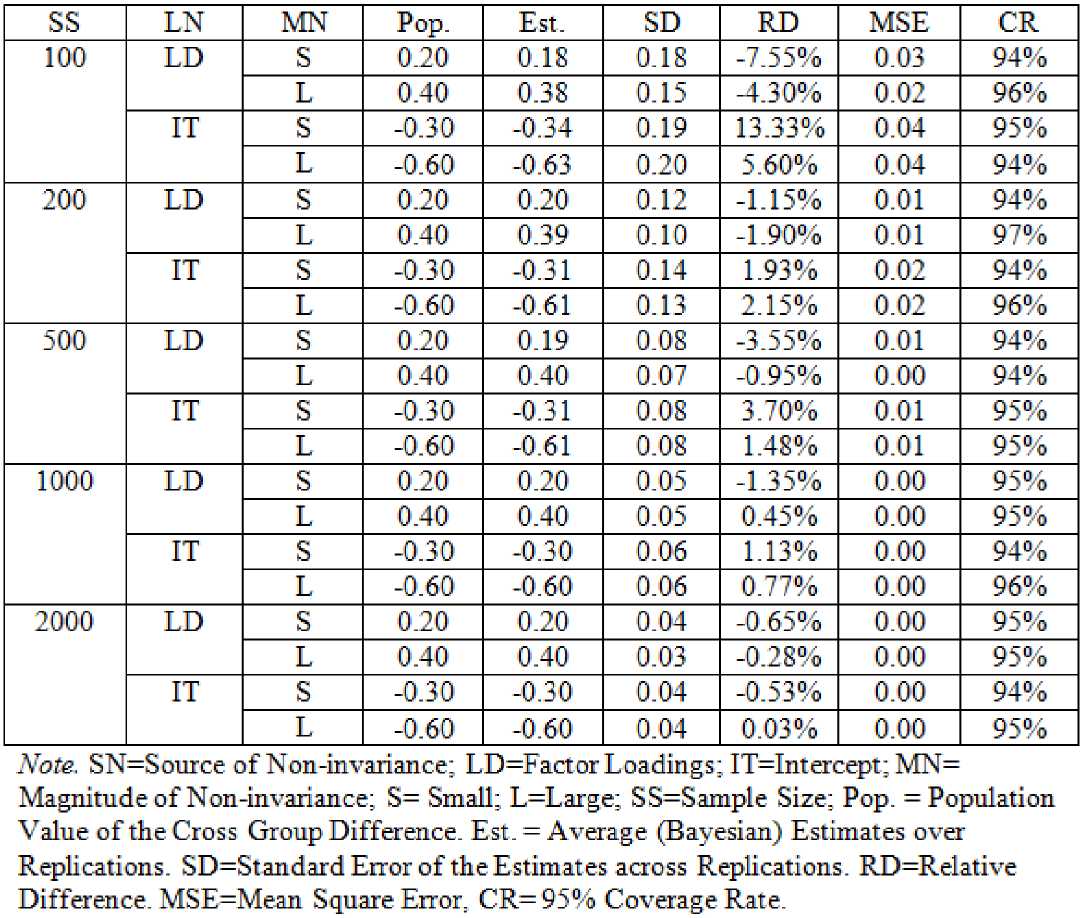
Table 5: Expected Difference of Observed Scores Conditional on Latent Trait (Epsilon =0) at Item Level (EDOI)



SS	LN	MN	Pop.	Est.	SD	RD	MSE	CR
100	LD	S	0.00	0.00	0.11	-	0.01	97%
		L	0.00	-0.01	0.11	-	0.01	95%
	IT	S	-0.30	-0.30	0.12	1.03%	0.01	97%
		L	-0.60	-0.60	0.13	-0.80%	0.02	94%
200	LD	S	0.00	0.00	0.08	-	0.01	95%
		L	0.00	0.00	0.07	-	0.01	96%
	IT	S	-0.30	-0.30	0.09	-0.53%	0.01	95%
		L	-0.60	-0.60	0.08	-0.28%	0.01	96%
500	LD	S	0.00	0.00	0.05	-	0.00	95%
		L	0.00	0.00	0.04	-	0.00	95%
	IT	S	-0.30	-0.30	0.05	0.13%	0.00	96%
		L	-0.60	-0.60	0.05	-0.05%	0.00	95%
1000	LD	S	0.00	0.00	0.03	-	0.00	95%
		L	0.00	0.00	0.03	-	0.00	96%
	IT	S	-0.30	-0.30	0.04	0.10%	0.00	95%
		L	-0.60	-0.60	0.04	0.02%	0.00	95%
2000	LD	S	0.00	0.00	0.02	-	0.00	96%
		L	0.00	0.00	0.02	-	0.00	95%
	IT	S	-0.30	-0.30	0.03	-0.13%	0.00	94%
		L	-0.60	-0.60	0.03	0.05%	0.00	95%

Note. SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; Pop. = Population Value of the Cross Group Difference. Est. = Average (Bayesian) Estimates over Replications. SD=Standard Error of the Estimates across Replications. RD=Relative Difference. MSE=Mean Square Error, CR=95% Coverage Rate.

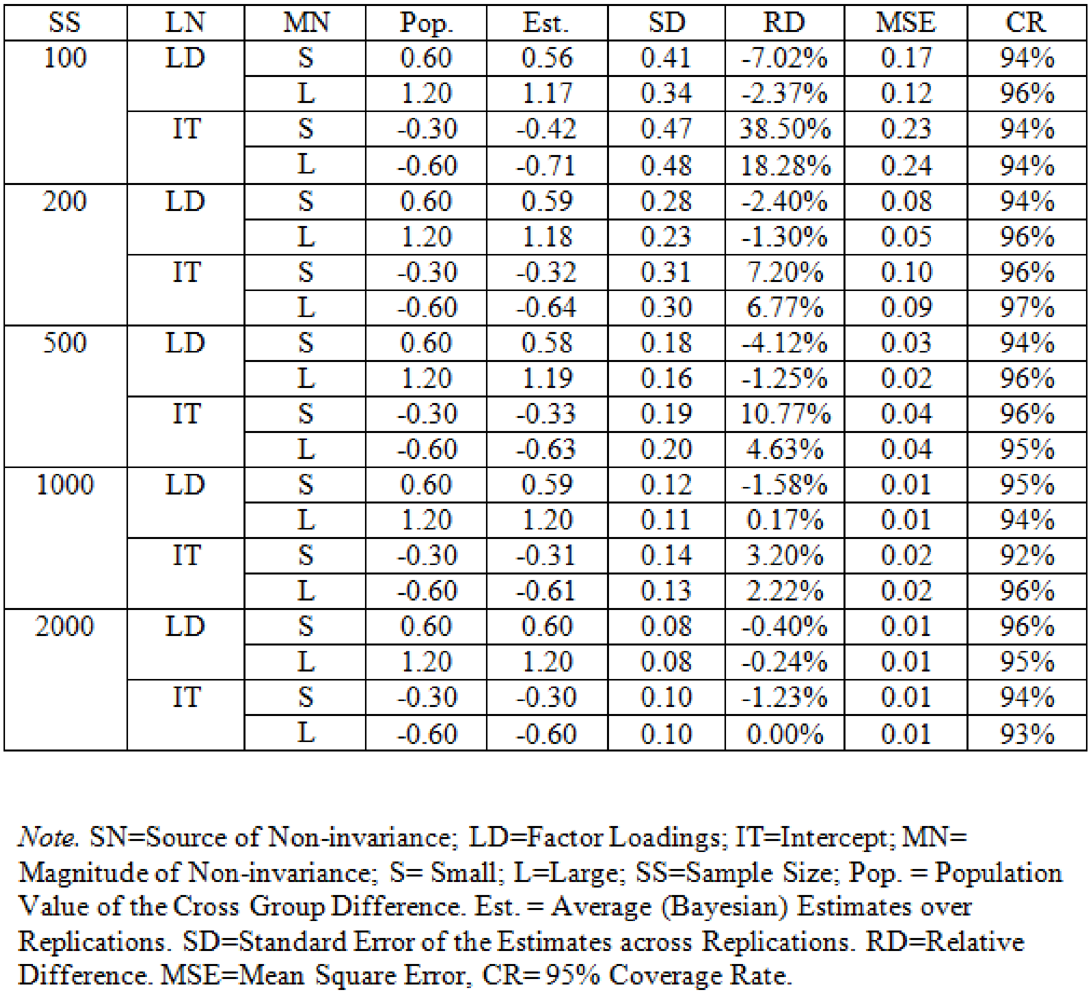
Table 6: Expected Difference of Observed Scores Conditional on Latent Trait (Epsilon=1) at Item Level (EDOI)



SS	LN	MN	Pop.	Est.	SD	RD	MSE	CR
100	LD	S	0.20	0.18	0.18	-7.55%	0.03	94%
		L	0.40	0.38	0.15	-4.30%	0.02	96%
	IT	S	-0.30	-0.34	0.19	13.33%	0.04	95%
		L	-0.60	-0.63	0.20	5.60%	0.04	94%
200	LD	S	0.20	0.20	0.12	-1.15%	0.01	94%
		L	0.40	0.39	0.10	-1.90%	0.01	97%
	IT	S	-0.30	-0.31	0.14	1.93%	0.02	94%
		L	-0.60	-0.61	0.13	2.15%	0.02	96%
500	LD	S	0.20	0.19	0.08	-3.55%	0.01	94%
		L	0.40	0.40	0.07	-0.95%	0.00	94%
	IT	S	-0.30	-0.31	0.08	3.70%	0.01	95%
		L	-0.60	-0.61	0.08	1.48%	0.01	95%
1000	LD	S	0.20	0.20	0.05	-1.35%	0.00	95%
		L	0.40	0.40	0.05	0.45%	0.00	95%
	IT	S	-0.30	-0.30	0.06	1.13%	0.00	94%
		L	-0.60	-0.60	0.06	0.77%	0.00	96%
2000	LD	S	0.20	0.20	0.04	-0.65%	0.00	95%
		L	0.40	0.40	0.03	-0.28%	0.00	95%
	IT	S	-0.30	-0.30	0.04	-0.53%	0.00	94%
		L	-0.60	-0.60	0.04	0.03%	0.00	95%

Note. SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; Pop. = Population Value of the Cross Group Difference. Est. = Average (Bayesian) Estimates over Replications. SD=Standard Error of the Estimates across Replications. RD=Relative Difference. MSE=Mean Square Error, CR= 95% Coverage Rate.

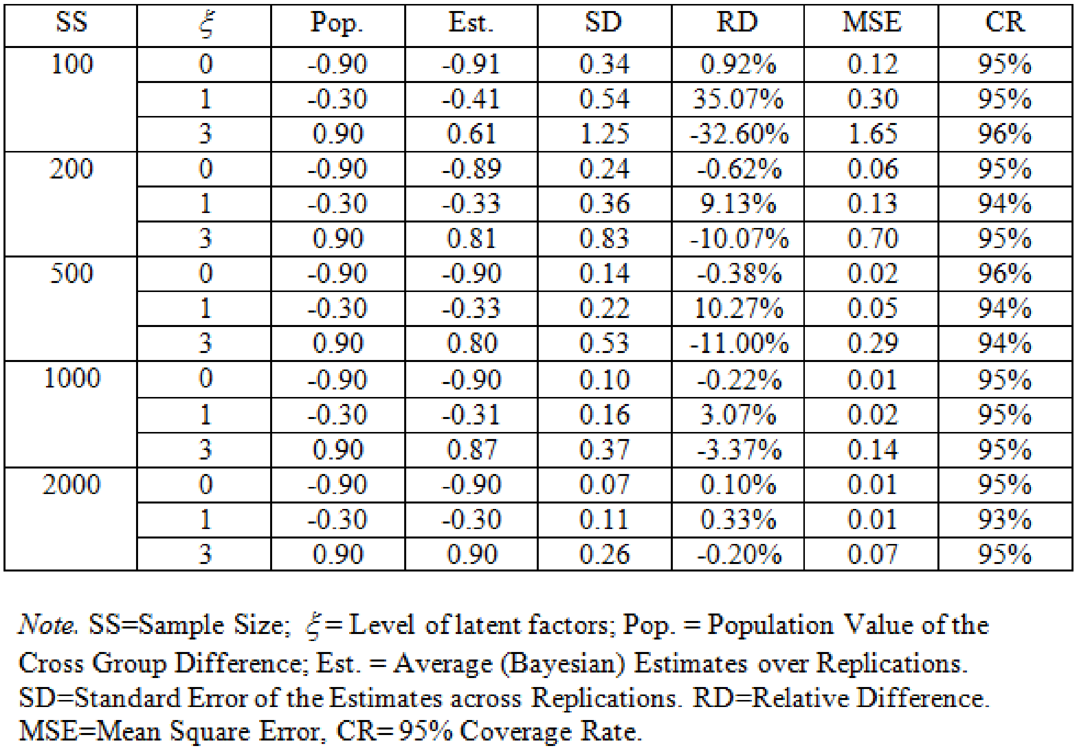
Table 7: Expected Difference of Observed Scores Conditional on Latent Trait (Epsilon =3) at Item Level (EDOI)



SS	LN	MN	Pop.	Est.	SD	RD	MSE	CR
100	LD	S	0.60	0.56	0.41	-7.02%	0.17	94%
		L	1.20	1.17	0.34	-2.37%	0.12	96%
	IT	S	-0.30	-0.42	0.47	38.50%	0.23	94%
		L	-0.60	-0.71	0.48	18.28%	0.24	94%
200	LD	S	0.60	0.59	0.28	-2.40%	0.08	94%
		L	1.20	1.18	0.23	-1.30%	0.05	96%
	IT	S	-0.30	-0.32	0.31	7.20%	0.10	96%
		L	-0.60	-0.64	0.30	6.77%	0.09	97%
500	LD	S	0.60	0.58	0.18	-4.12%	0.03	94%
		L	1.20	1.19	0.16	-1.25%	0.02	96%
	IT	S	-0.30	-0.33	0.19	10.77%	0.04	96%
		L	-0.60	-0.63	0.20	4.63%	0.04	95%
1000	LD	S	0.60	0.59	0.12	-1.58%	0.01	95%
		L	1.20	1.20	0.11	0.17%	0.01	94%
	IT	S	-0.30	-0.31	0.14	3.20%	0.02	92%
		L	-0.60	-0.61	0.13	2.22%	0.02	96%
2000	LD	S	0.60	0.60	0.08	-0.40%	0.01	96%
		L	1.20	1.20	0.08	-0.24%	0.01	95%
	IT	S	-0.30	-0.30	0.10	-1.23%	0.01	94%
		L	-0.60	-0.60	0.10	0.00%	0.01	93%

Note. SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; Pop. = Population Value of the Cross Group Difference. Est. = Average (Bayesian) Estimates over Replications. SD=Standard Error of the Estimates across Replications. RD=Relative Difference. MSE=Mean Square Error, CR= 95% Coverage Rate.

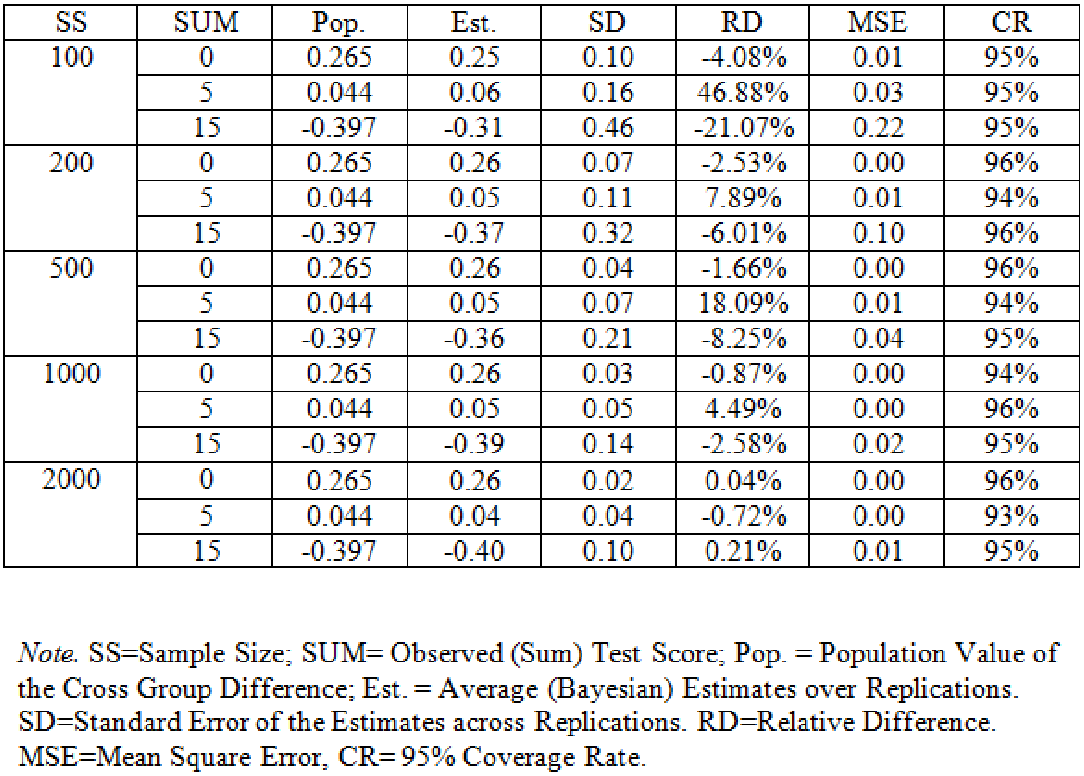
Table 8: Expected Difference of Observed Scores Conditional on Latent Traits at Test (Sum Score) Level (EDOT)



SS	ξ	Pop.	Est.	SD	RD	MSE	CR
100	0	-0.90	-0.91	0.34	0.92%	0.12	95%
	1	-0.30	-0.41	0.54	35.07%	0.30	95%
	3	0.90	0.61	1.25	-32.60%	1.65	96%
200	0	-0.90	-0.89	0.24	-0.62%	0.06	95%
	1	-0.30	-0.33	0.36	9.13%	0.13	94%
	3	0.90	0.81	0.83	-10.07%	0.70	95%
500	0	-0.90	-0.90	0.14	-0.38%	0.02	96%
	1	-0.30	-0.33	0.22	10.27%	0.05	94%
	3	0.90	0.80	0.53	-11.00%	0.29	94%
1000	0	-0.90	-0.90	0.10	-0.22%	0.01	95%
	1	-0.30	-0.31	0.16	3.07%	0.02	95%
	3	0.90	0.87	0.37	-3.37%	0.14	95%
2000	0	-0.90	-0.90	0.07	0.10%	0.01	95%
	1	-0.30	-0.30	0.11	0.33%	0.01	93%
	3	0.90	0.90	0.26	-0.20%	0.07	95%

Note. SS=Sample Size; ξ= Level of latent factors; Pop. = Population Value of the Cross Group Difference; Est. = Average (Bayesian) Estimates over Replications. SD=Standard Error of the Estimates across Replications. RD=Relative Difference. MSE=Mean Square Error, CR= 95% Coverage Rate.

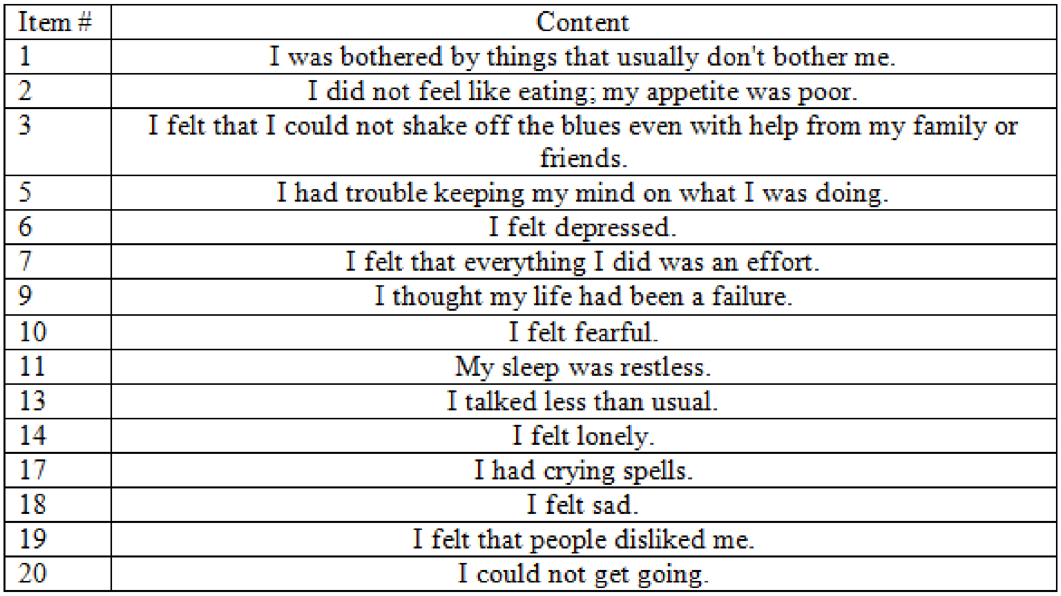
Table 9: Expected Difference of Latent Traits Conditional on Observed (Sum) Scores (EDLT)



SS	SUM	Pop.	Est.	SD	RD	MSE	CR
100	0	0.265	0.25	0.10	-4.08%	0.01	95%
	5	0.044	0.06	0.16	46.88%	0.03	95%
	15	-0.397	-0.31	0.46	-21.07%	0.22	95%
200	0	0.265	0.26	0.07	-2.53%	0.00	96%
	5	0.044	0.05	0.11	7.89%	0.01	94%
	15	-0.397	-0.37	0.32	-6.01%	0.10	96%
500	0	0.265	0.26	0.04	-1.66%	0.00	96%
	5	0.044	0.05	0.07	18.09%	0.01	94%
	15	-0.397	-0.36	0.21	-8.25%	0.04	95%
1000	0	0.265	0.26	0.03	-0.87%	0.00	94%
	5	0.044	0.05	0.05	4.49%	0.00	96%
	15	-0.397	-0.39	0.14	-2.58%	0.02	95%
2000	0	0.265	0.26	0.02	0.04%	0.00	96%
	5	0.044	0.04	0.04	-0.72%	0.00	93%
	15	-0.397	-0.40	0.10	0.21%	0.01	95%

Note. SS=Sample Size; SUM= Observed (Sum) Test Score; Pop. = Population Value of the Cross Group Difference; Est. = Average (Bayesian) Estimates over Replications. SD=Standard Error of the Estimates across Replications. RD=Relative Difference. MSE=Mean Square Error, CR= 95% Coverage Rate.

Table 10: Content of the Shortened Version of the CES-D Scale (15 Items)

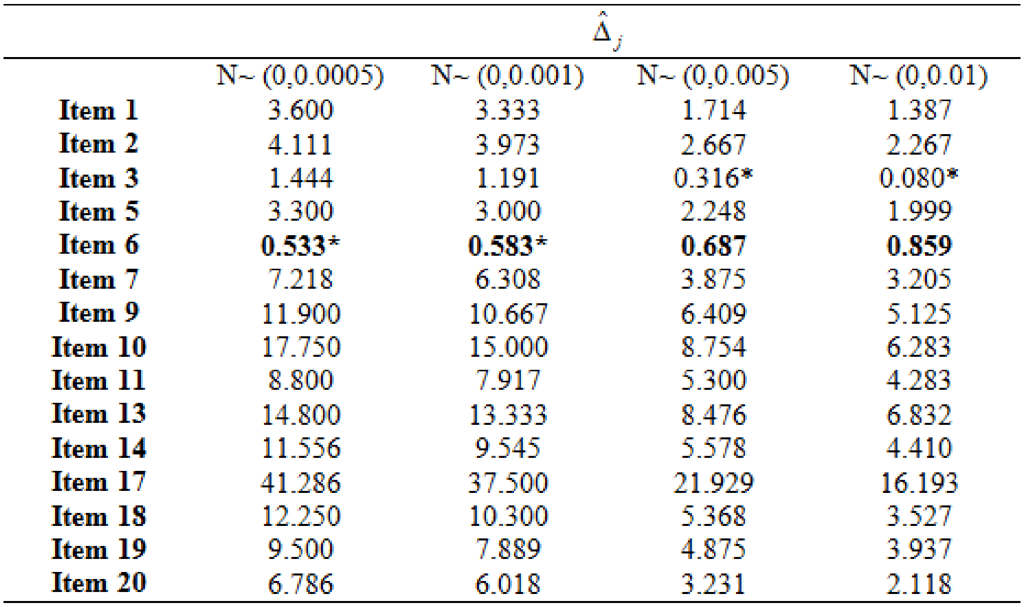


Item #	Content
1	I was bothered by things that usually don't bother me.
2	I did not feel like eating; my appetite was poor.
3	I felt that I could not shake off the blues even with help from my family or friends.
5	I had trouble keeping my mind on what I was doing.
6	I felt depressed.
7	I felt that everything I did was an effort.
9	I thought my life had been a failure.
10	I felt fearful.
11	My sleep was restless.
13	I talked less than usual.
14	I felt lonely.
17	I had crying spells.
18	I felt sad.
19	I felt that people disliked me.
20	I could not get going.

Table 11: Descriptive Statistics and Correlation Coefficients between Items

	X1	X3	X4	X5	X6	X7	X9	X10	X11	X13	X14	X17	X18	X19	X20
X1	0.781														
X3	0.359	0.717													
X4	0.399	0.379	0.670												
X5	0.293	0.285	0.321	0.764											
X6	0.436	0.378	0.450	0.441	0.716										
X7	0.313	0.320	0.357	0.403	0.483	0.790									
X9	0.299	0.251	0.326	0.310	0.406	0.400	0.744								
X10	0.295	0.280	0.331	0.289	0.374	0.343	0.382	0.588							
X11	0.280	0.324	0.272	0.279	0.339	0.306	0.282	0.296	0.834						
X13	0.256	0.254	0.281	0.273	0.335	0.301	0.296	0.260	0.246	0.792					
X14	0.316	0.282	0.349	0.295	0.415	0.353	0.382	0.391	0.285	0.369	0.673				
X17	0.336	0.289	0.345	0.262	0.387	0.293	0.304	0.402	0.276	0.254	0.393	0.583			
X18	0.380	0.313	0.393	0.299	0.458	0.356	0.377	0.414	0.314	0.296	0.459	0.621	0.617		
X19	0.281	0.249	0.306	0.261	0.348	0.306	0.344	0.365	0.238	0.268	0.387	0.390	0.462	0.542	
X20	0.258	0.259	0.332	0.263	0.332	0.327	0.342	0.362	0.241	0.243	0.396	0.382	0.431	0.395	0.480
Mean	0.759	0.553	0.405	0.613	0.621	0.607	0.471	0.299	0.656	0.585	0.375	0.330	0.404	0.295	0.157

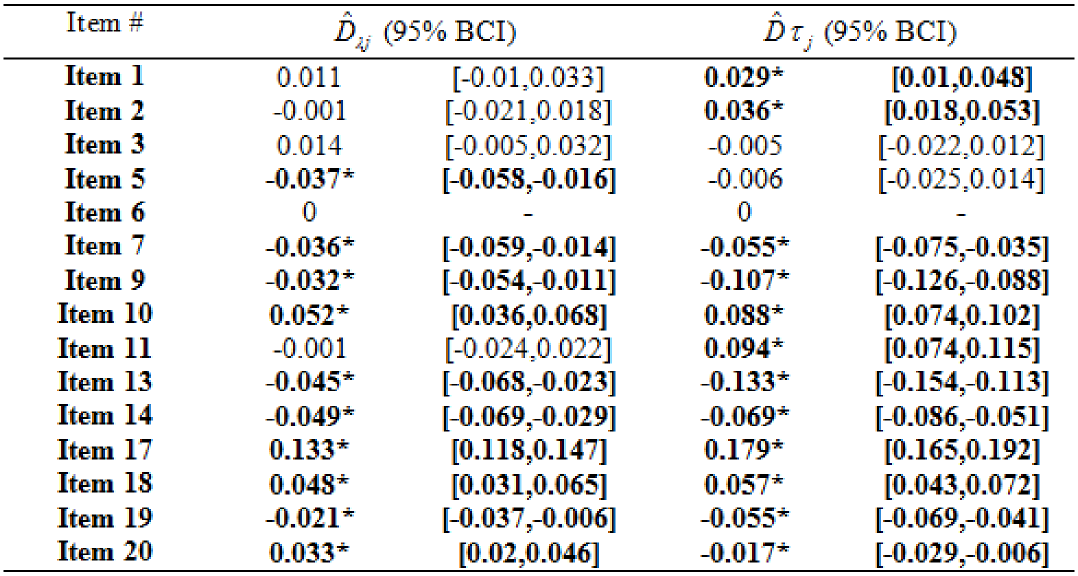
Table 12: Values of Selection Index in Selecting RI in the Empirical Analysis



	$\hat{\Delta}_j$			
	N~ (0,0.0005)	N~ (0,0.001)	N~ (0,0.005)	N~ (0,0.01)
Item 1	3.600	3.333	1.714	1.387
Item 2	4.111	3.973	2.667	2.267
Item 3	1.444	1.191	0.316*	0.080*
Item 5	3.300	3.000	2.248	1.999
Item 6	**0.533***	**0.583***	**0.687**	**0.859**
Item 7	7.218	6.308	3.875	3.205
Item 9	11.900	10.667	6.409	5.125
Item 10	17.750	15.000	8.754	6.283
Item 11	8.800	7.917	5.300	4.283
Item 13	14.800	13.333	8.476	6.832
Item 14	11.556	9.545	5.578	4.410
Item 17	41.286	37.500	21.929	16.193
Item 18	12.250	10.300	5.368	3.527
Item 19	9.500	7.889	4.875	3.937
Item 20	6.786	6.018	3.231	2.118

Note. $\hat{\Delta}_j$ is the section index; asterisk in each column indicates the item associated with the smallest value of $\hat{\Delta}_j$.

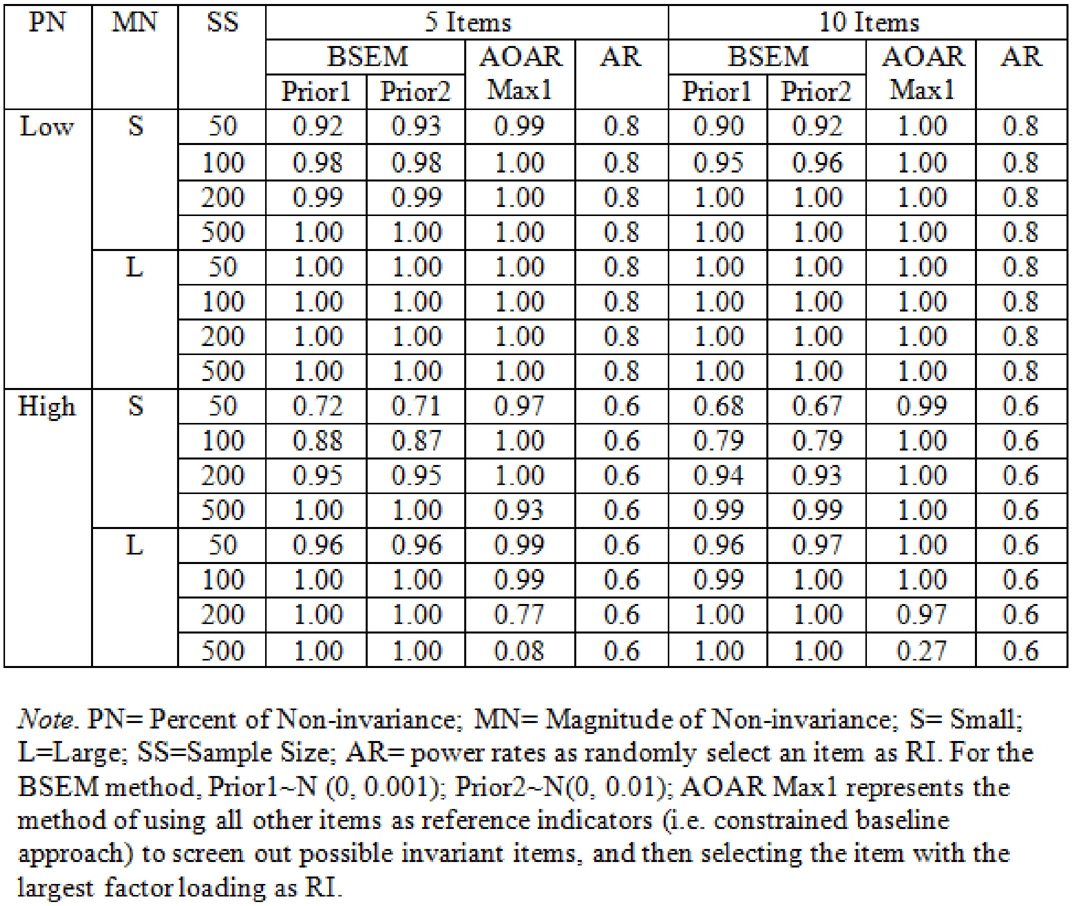
Table 13: Locating Non-invariant Parameters in the Empirical Analysis



Item #	$\hat{D}_{\lambda j}$ (95% BCI)		$\hat{D}\tau_j$ (95% BCI)	
Item 1	0.011	[-0.01,0.033]	**0.029***	**[0.01,0.048]**
Item 2	-0.001	[-0.021,0.018]	**0.036***	**[0.018,0.053]**
Item 3	0.014	[-0.005,0.032]	-0.005	[-0.022,0.012]
Item 5	**-0.037***	**[-0.058,-0.016]**	-0.006	[-0.025,0.014]
Item 6	0	-	0	-
Item 7	**-0.036***	**[-0.059,-0.014]**	**-0.055***	**[-0.075,-0.035]**
Item 9	**-0.032***	**[-0.054,-0.011]**	**-0.107***	**[-0.126,-0.088]**
Item 10	**0.052***	**[0.036,0.068]**	**0.088***	**[0.074,0.102]**
Item 11	-0.001	[-0.024,0.022]	**0.094***	**[0.074,0.115]**
Item 13	**-0.045***	**[-0.068,-0.023]**	**-0.133***	**[-0.154,-0.113]**
Item 14	**-0.049***	**[-0.069,-0.029]**	**-0.069***	**[-0.086,-0.051]**
Item 17	**0.133***	**[0.118,0.147]**	**0.179***	**[0.165,0.192]**
Item 18	**0.048***	**[0.031,0.065]**	**0.057***	**[0.043,0.072]**
Item 19	**-0.021***	**[-0.037,-0.006]**	**-0.055***	**[-0.069,-0.041]**
Item 20	**0.033***	**[0.02,0.046]**	**-0.017***	**[-0.029,-0.006]**

Note. $\hat{D}_{\lambda j}$ and $\hat{D}\tau_j$ indicate the point estimate of the cross-group difference for the factor loadings and intercepts, respectively; BCI = Bayesian Credible Interval (BCI); asterisks indicate the identified non-invariant parameters.

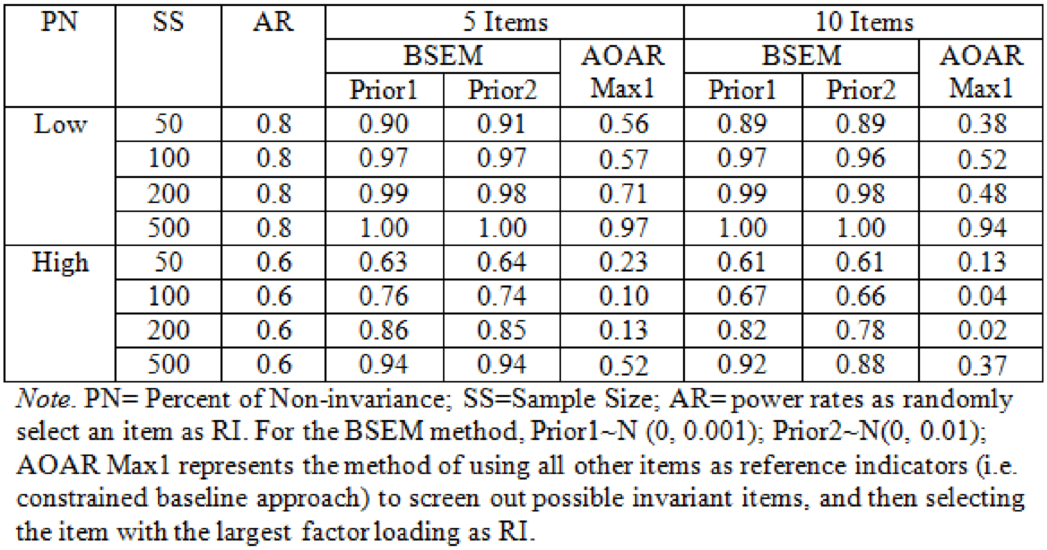
Table 14: Power Rates of Selecting a Reference Indicator with Invariant Factor Loading Only.



PN	MN	SS	5 Items				10 Items			
			BSEM		AOAR	AR	BSEM		AOAR	AR
			Prior1	Prior2	Max1		Prior1	Prior2	Max1	
Low	S	50	0.92	0.93	0.99	0.8	0.90	0.92	1.00	0.8
		100	0.98	0.98	1.00	0.8	0.95	0.96	1.00	0.8
		200	0.99	0.99	1.00	0.8	1.00	1.00	1.00	0.8
		500	1.00	1.00	1.00	0.8	1.00	1.00	1.00	0.8
	L	50	1.00	1.00	1.00	0.8	1.00	1.00	1.00	0.8
		100	1.00	1.00	1.00	0.8	1.00	1.00	1.00	0.8
		200	1.00	1.00	1.00	0.8	1.00	1.00	1.00	0.8
		500	1.00	1.00	1.00	0.8	1.00	1.00	1.00	0.8
High	S	50	0.72	0.71	0.97	0.6	0.68	0.67	0.99	0.6
		100	0.88	0.87	1.00	0.6	0.79	0.79	1.00	0.6
		200	0.95	0.95	1.00	0.6	0.94	0.93	1.00	0.6
		500	1.00	1.00	0.93	0.6	0.99	0.99	1.00	0.6
	L	50	0.96	0.96	0.99	0.6	0.96	0.97	1.00	0.6
		100	1.00	1.00	0.99	0.6	0.99	1.00	1.00	0.6
		200	1.00	1.00	0.77	0.6	1.00	1.00	0.97	0.6
		500	1.00	1.00	0.08	0.6	1.00	1.00	0.27	0.6

Note. PN= Percent of Non-invariance; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select an item as RI. For the BSEM method, Prior1~N (0, 0.001); Prior2~N(0, 0.01); AOAR Max1 represents the method of using all other items as reference indicators (i.e. constrained baseline approach) to screen out possible invariant items, and then selecting the item with the largest factor loading as RI.

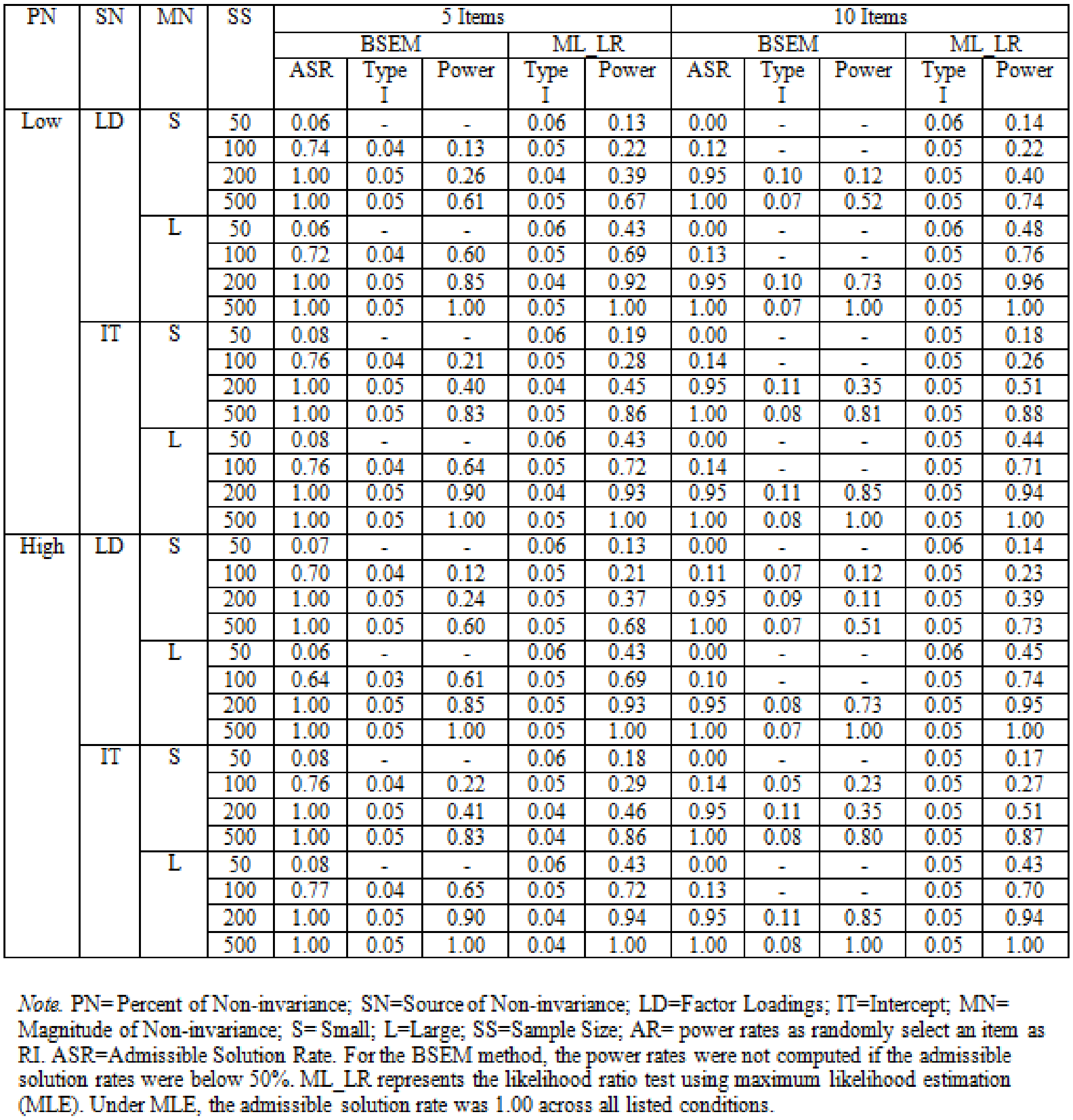
Table 15: Power Rates of Selecting a Reference Indicator



PN	SS	AR	5 Items			10 Items		
			BSEM		AOAR	BSEM		AOAR
			Prior1	Prior2	Max1	Prior1	Prior2	Max1
Low	50	0.8	0.90	0.91	0.56	0.89	0.89	0.38
	100	0.8	0.97	0.97	0.57	0.97	0.96	0.52
	200	0.8	0.99	0.98	0.71	0.99	0.98	0.48
	500	0.8	1.00	1.00	0.97	1.00	1.00	0.94
High	50	0.6	0.63	0.64	0.23	0.61	0.61	0.13
	100	0.6	0.76	0.74	0.10	0.67	0.66	0.04
	200	0.6	0.86	0.85	0.13	0.82	0.78	0.02
	500	0.6	0.94	0.94	0.52	0.92	0.88	0.37

Note. PN= Percent of Non-invariance; SS=Sample Size; AR= power rates as randomly select an item as RI. For the BSEM method, Prior1~N (0, 0.001); Prior2~N(0, 0.01); AOAR Max1 represents the method of using all other items as reference indicators (i.e. constrained baseline approach) to screen out possible invariant items, and then selecting the item with the largest factor loading as RI.

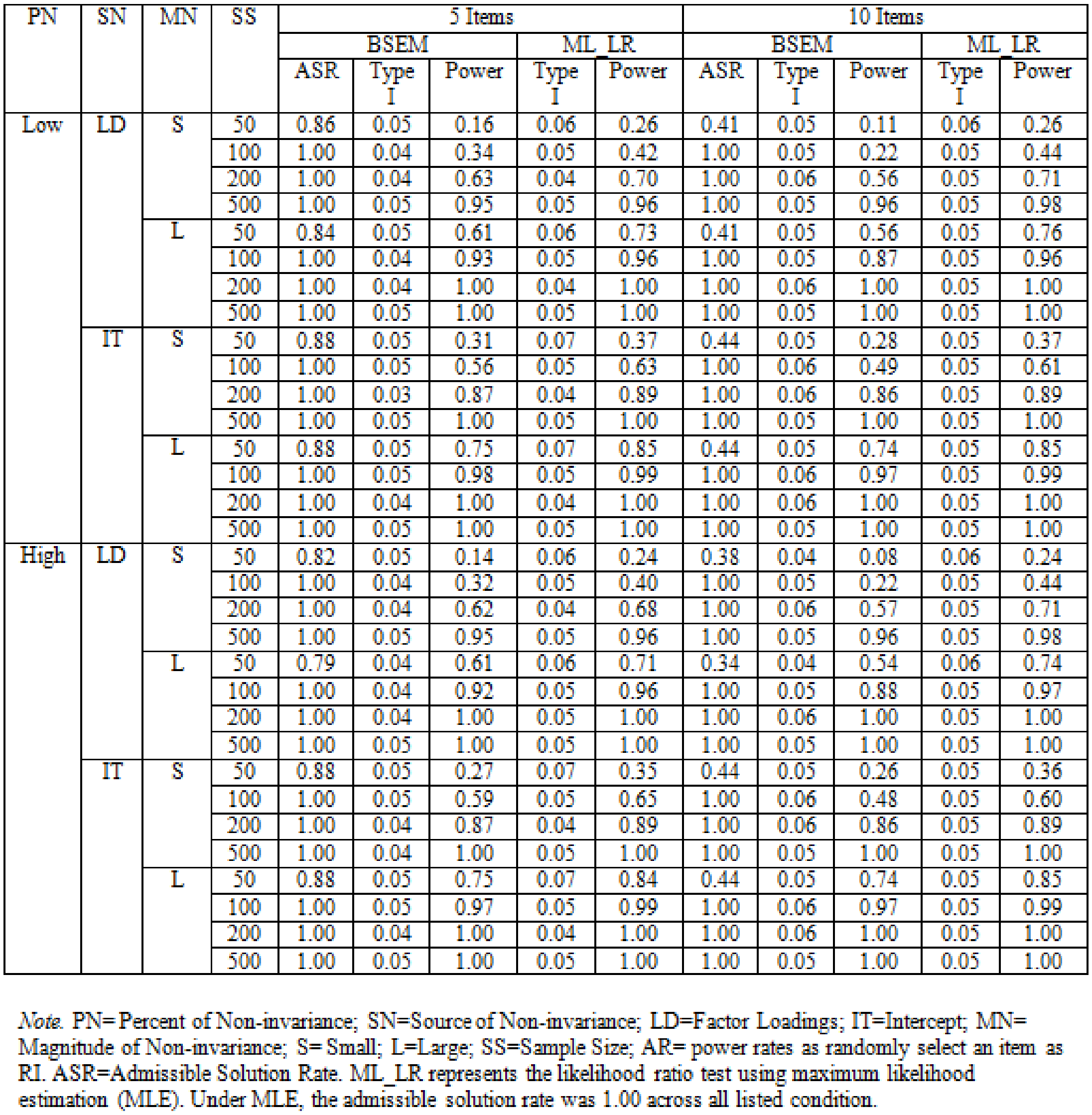
Table 16: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (1 RI with Factor Loading of 0.4)



PN	SN	MN	SS	5 Items					10 Items				
				BSEM			ML_LR		BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power	ASR	Type I	Power	Type I	Power
Low	LD	S	50	0.06	-	-	0.06	0.13	0.00	-	-	0.06	0.14
			100	0.74	0.04	0.13	0.05	0.22	0.12	-	-	0.05	0.22
			200	1.00	0.05	0.26	0.04	0.39	0.95	0.10	0.12	0.05	0.40
			500	1.00	0.05	0.61	0.05	0.67	1.00	0.07	0.52	0.05	0.74
		L	50	0.06	-	-	0.06	0.43	0.00	-	-	0.06	0.48
			100	0.72	0.04	0.60	0.05	0.69	0.13	-	-	0.05	0.76
			200	1.00	0.05	0.85	0.04	0.92	0.95	0.10	0.73	0.05	0.96
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.07	1.00	0.05	1.00
	IT	S	50	0.08	-	-	0.06	0.19	0.00	-	-	0.05	0.18
			100	0.76	0.04	0.21	0.05	0.28	0.14	-	-	0.05	0.26
			200	1.00	0.05	0.40	0.04	0.45	0.95	0.11	0.35	0.05	0.51
			500	1.00	0.05	0.83	0.05	0.86	1.00	0.08	0.81	0.05	0.88
		L	50	0.08	-	-	0.06	0.43	0.00	-	-	0.05	0.44
			100	0.76	0.04	0.64	0.05	0.72	0.14	-	-	0.05	0.71
			200	1.00	0.05	0.90	0.04	0.93	0.95	0.11	0.85	0.05	0.94
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.08	1.00	0.05	1.00
High	LD	S	50	0.07	-	-	0.06	0.13	0.00	-	-	0.06	0.14
			100	0.70	0.04	0.12	0.05	0.21	0.11	0.07	0.12	0.05	0.23
			200	1.00	0.05	0.24	0.05	0.37	0.95	0.09	0.11	0.05	0.39
			500	1.00	0.05	0.60	0.05	0.68	1.00	0.07	0.51	0.05	0.73
		L	50	0.06	-	-	0.06	0.43	0.00	-	-	0.06	0.45
			100	0.64	0.03	0.61	0.05	0.69	0.10	-	-	0.05	0.74
			200	1.00	0.05	0.85	0.05	0.93	0.95	0.08	0.73	0.05	0.95
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.07	1.00	0.05	1.00
	IT	S	50	0.08	-	-	0.06	0.18	0.00	-	-	0.05	0.17
			100	0.76	0.04	0.22	0.05	0.29	0.14	0.05	0.23	0.05	0.27
			200	1.00	0.05	0.41	0.04	0.46	0.95	0.11	0.35	0.05	0.51
			500	1.00	0.05	0.83	0.04	0.86	1.00	0.08	0.80	0.05	0.87
		L	50	0.08	-	-	0.06	0.43	0.00	-	-	0.05	0.43
			100	0.77	0.04	0.65	0.05	0.72	0.13	-	-	0.05	0.70
			200	1.00	0.05	0.90	0.04	0.94	0.95	0.11	0.85	0.05	0.94
			500	1.00	0.05	1.00	0.04	1.00	1.00	0.08	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select an item as RI. ASR=Admissible Solution Rate. For the BSEM method, the power rates were not computed if the admissible solution rates were below 50%. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed conditions.

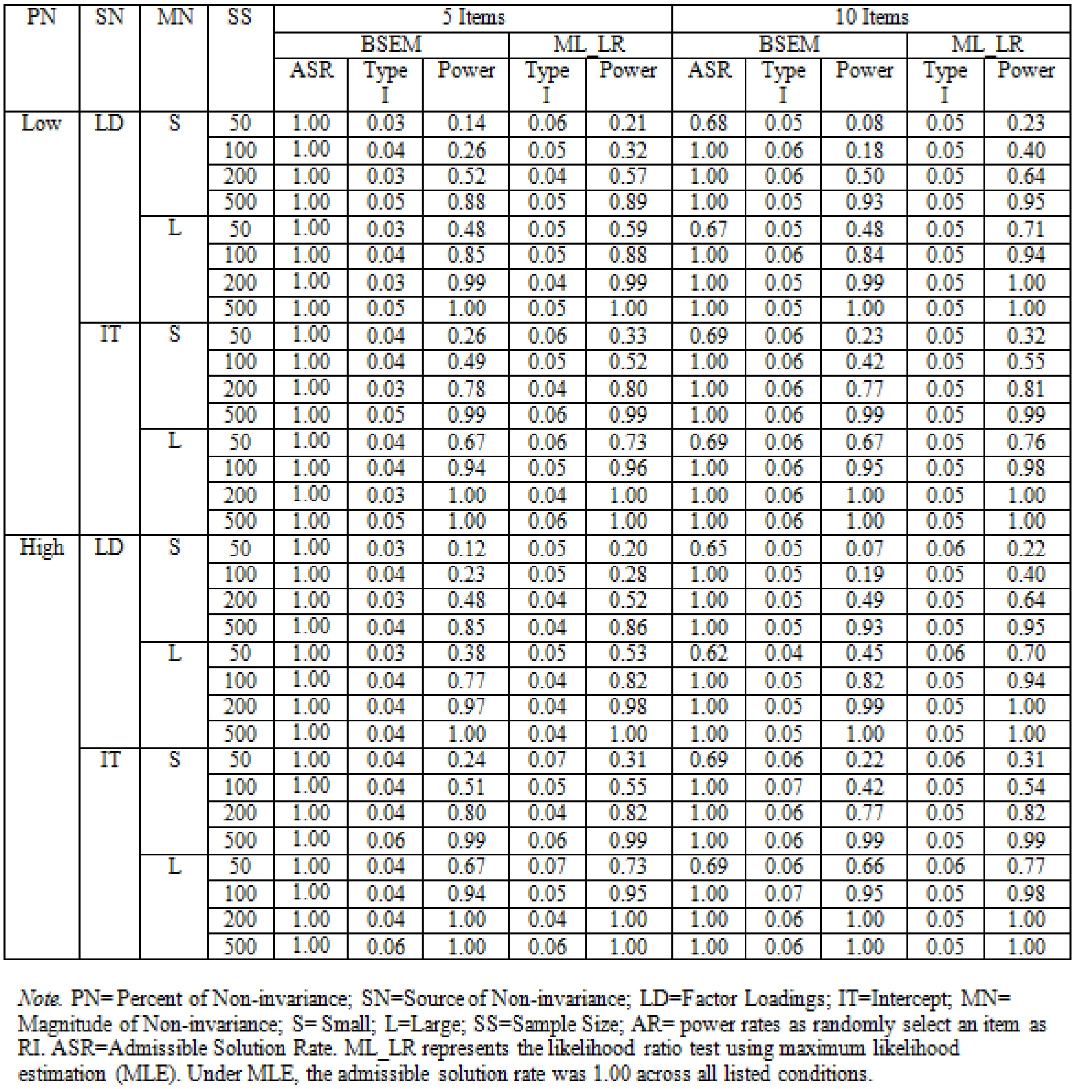
Table 17: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (1 RI with Factor Loading of 0.8)



PN	SN	MN	SS	5 Items					10 Items				
				BSEM			ML_LR		BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power	ASR	Type I	Power	Type I	Power
Low	LD	S	50	0.86	0.05	0.16	0.06	0.26	0.41	0.05	0.11	0.06	0.26
			100	1.00	0.04	0.34	0.05	0.42	1.00	0.05	0.22	0.05	0.44
			200	1.00	0.04	0.63	0.04	0.70	1.00	0.06	0.56	0.05	0.71
			500	1.00	0.05	0.95	0.05	0.96	1.00	0.05	0.96	0.05	0.98
		L	50	0.84	0.05	0.61	0.06	0.73	0.41	0.05	0.56	0.05	0.76
			100	1.00	0.04	0.93	0.05	0.96	1.00	0.05	0.87	0.05	0.96
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.88	0.05	0.31	0.07	0.37	0.44	0.05	0.28	0.05	0.37
			100	1.00	0.05	0.56	0.05	0.63	1.00	0.06	0.49	0.05	0.61
			200	1.00	0.03	0.87	0.04	0.89	1.00	0.06	0.86	0.05	0.89
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
		L	50	0.88	0.05	0.75	0.07	0.85	0.44	0.05	0.74	0.05	0.85
			100	1.00	0.05	0.98	0.05	0.99	1.00	0.06	0.97	0.05	0.99
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
High	LD	S	50	0.82	0.05	0.14	0.06	0.24	0.38	0.04	0.08	0.06	0.24
			100	1.00	0.04	0.32	0.05	0.40	1.00	0.05	0.22	0.05	0.44
			200	1.00	0.04	0.62	0.04	0.68	1.00	0.06	0.57	0.05	0.71
			500	1.00	0.05	0.95	0.05	0.96	1.00	0.05	0.96	0.05	0.98
		L	50	0.79	0.04	0.61	0.06	0.71	0.34	0.04	0.54	0.06	0.74
			100	1.00	0.04	0.92	0.05	0.96	1.00	0.05	0.88	0.05	0.97
			200	1.00	0.04	1.00	0.05	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	0.88	0.05	0.27	0.07	0.35	0.44	0.05	0.26	0.05	0.36
			100	1.00	0.05	0.59	0.05	0.65	1.00	0.06	0.48	0.05	0.60
			200	1.00	0.04	0.87	0.04	0.89	1.00	0.06	0.86	0.05	0.89
			500	1.00	0.04	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
		L	50	0.88	0.05	0.75	0.07	0.84	0.44	0.05	0.74	0.05	0.85
			100	1.00	0.05	0.97	0.05	0.99	1.00	0.06	0.97	0.05	0.99
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select an item as RI. ASR=Admissible Solution Rate. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed condition.

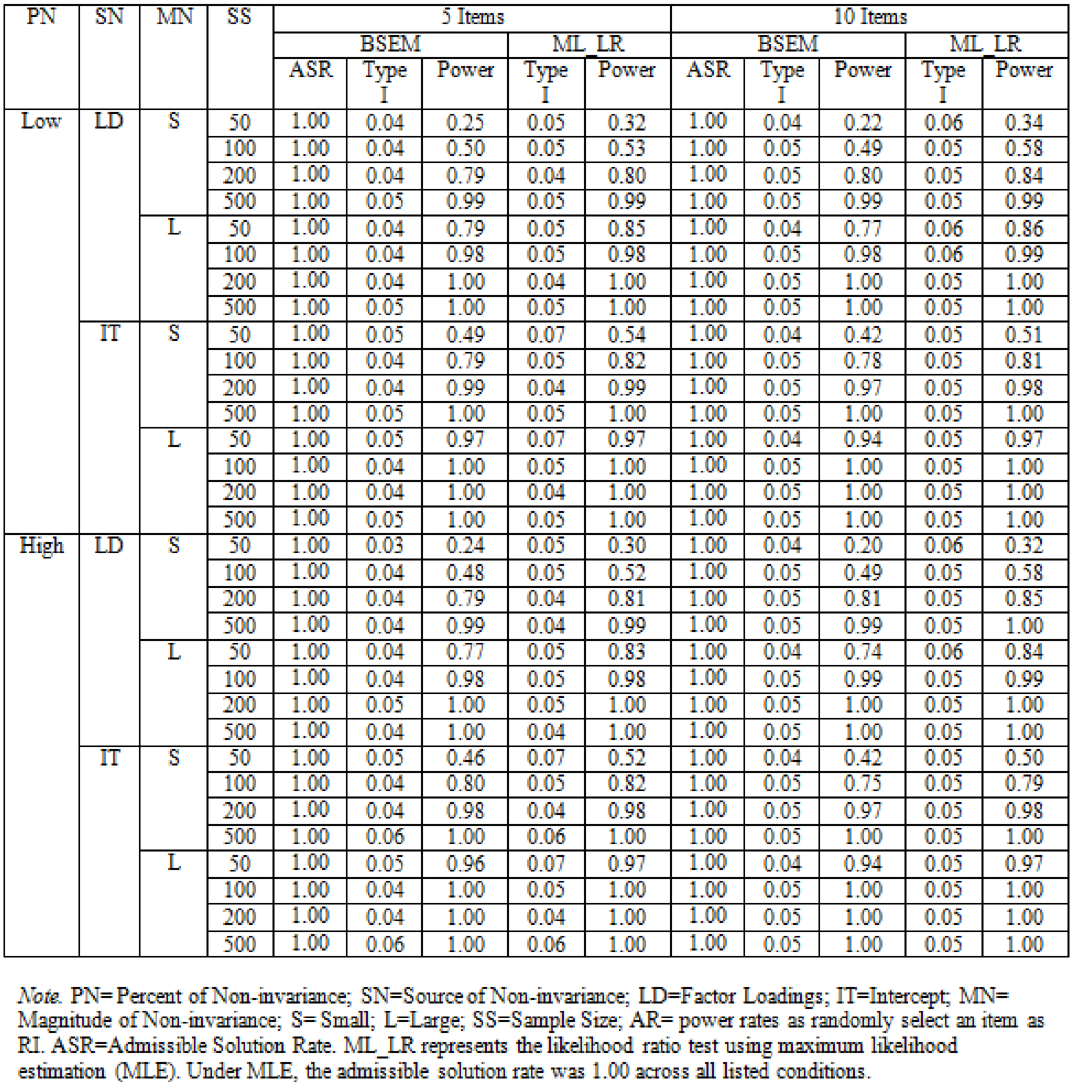
Table 18: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (3 RIs with Factor Loadings of 0.4)



PN	SN	MN	SS	5 Items					10 Items				
				BSEM			ML_LR		BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power	ASR	Type I	Power	Type I	Power
Low	LD	S	50	1.00	0.03	0.14	0.06	0.21	0.68	0.05	0.08	0.05	0.23
			100	1.00	0.04	0.26	0.05	0.32	1.00	0.06	0.18	0.05	0.40
			200	1.00	0.03	0.52	0.04	0.57	1.00	0.06	0.50	0.05	0.64
			500	1.00	0.05	0.88	0.05	0.89	1.00	0.05	0.93	0.05	0.95
		L	50	1.00	0.03	0.48	0.05	0.59	0.67	0.05	0.48	0.05	0.71
			100	1.00	0.04	0.85	0.05	0.88	1.00	0.06	0.84	0.05	0.94
			200	1.00	0.03	0.99	0.04	0.99	1.00	0.05	0.99	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	1.00	0.04	0.26	0.06	0.33	0.69	0.06	0.23	0.05	0.32
			100	1.00	0.04	0.49	0.05	0.52	1.00	0.06	0.42	0.05	0.55
			200	1.00	0.03	0.78	0.04	0.80	1.00	0.06	0.77	0.05	0.81
			500	1.00	0.05	0.99	0.06	0.99	1.00	0.06	0.99	0.05	0.99
		L	50	1.00	0.04	0.67	0.06	0.73	0.69	0.06	0.67	0.05	0.76
			100	1.00	0.04	0.94	0.05	0.96	1.00	0.06	0.95	0.05	0.98
			200	1.00	0.03	1.00	0.04	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.06	1.00	1.00	0.06	1.00	0.05	1.00
High	LD	S	50	1.00	0.03	0.12	0.05	0.20	0.65	0.05	0.07	0.06	0.22
			100	1.00	0.04	0.23	0.05	0.28	1.00	0.05	0.19	0.05	0.40
			200	1.00	0.03	0.48	0.04	0.52	1.00	0.05	0.49	0.05	0.64
			500	1.00	0.04	0.85	0.04	0.86	1.00	0.05	0.93	0.05	0.95
		L	50	1.00	0.03	0.38	0.05	0.53	0.62	0.04	0.45	0.06	0.70
			100	1.00	0.04	0.77	0.04	0.82	1.00	0.05	0.82	0.05	0.94
			200	1.00	0.04	0.97	0.04	0.98	1.00	0.05	0.99	0.05	1.00
			500	1.00	0.04	1.00	0.04	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	1.00	0.04	0.24	0.07	0.31	0.69	0.06	0.22	0.06	0.31
			100	1.00	0.04	0.51	0.05	0.55	1.00	0.07	0.42	0.05	0.54
			200	1.00	0.04	0.80	0.04	0.82	1.00	0.06	0.77	0.05	0.82
			500	1.00	0.06	0.99	0.06	0.99	1.00	0.06	0.99	0.05	0.99
		L	50	1.00	0.04	0.67	0.07	0.73	0.69	0.06	0.66	0.06	0.77
			100	1.00	0.04	0.94	0.05	0.95	1.00	0.07	0.95	0.05	0.98
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.06	1.00	0.05	1.00
			500	1.00	0.06	1.00	0.06	1.00	1.00	0.06	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select an item as RI. ASR=Admissible Solution Rate. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed conditions.

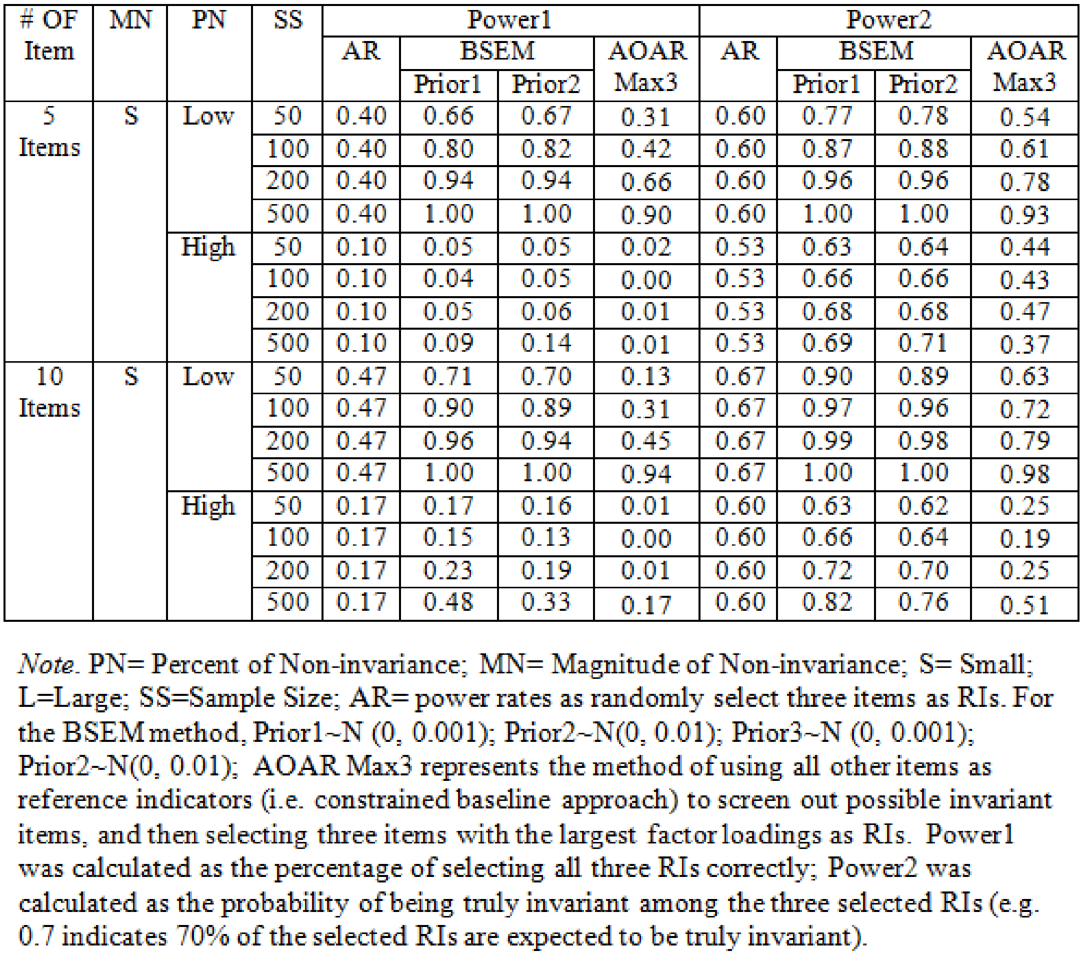
Table 19: Admissible Solution Rates, Type I Error Rates, and Power Rates in Locating Non-invariance (3 RIs with Factor Loadings of 0.8)



PN	SN	MN	SS	5 Items					10 Items				
				BSEM			ML_LR		BSEM			ML_LR	
				ASR	Type I	Power	Type I	Power	ASR	Type I	Power	Type I	Power
Low	LD	S	50	1.00	0.04	0.25	0.05	0.32	1.00	0.04	0.22	0.06	0.34
			100	1.00	0.04	0.50	0.05	0.53	1.00	0.05	0.49	0.05	0.58
			200	1.00	0.04	0.79	0.04	0.80	1.00	0.05	0.80	0.05	0.84
			500	1.00	0.05	0.99	0.05	0.99	1.00	0.05	0.99	0.05	0.99
		L	50	1.00	0.04	0.79	0.05	0.85	1.00	0.04	0.77	0.06	0.86
			100	1.00	0.04	0.98	0.05	0.98	1.00	0.05	0.98	0.06	0.99
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.05	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	1.00	0.05	0.49	0.07	0.54	1.00	0.04	0.42	0.05	0.51
			100	1.00	0.04	0.79	0.05	0.82	1.00	0.05	0.78	0.05	0.81
			200	1.00	0.04	0.99	0.04	0.99	1.00	0.05	0.97	0.05	0.98
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
		L	50	1.00	0.05	0.97	0.07	0.97	1.00	0.04	0.94	0.05	0.97
			100	1.00	0.04	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.05	1.00	0.05	1.00
			500	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
High	LD	S	50	1.00	0.03	0.24	0.05	0.30	1.00	0.04	0.20	0.06	0.32
			100	1.00	0.04	0.48	0.05	0.52	1.00	0.05	0.49	0.05	0.58
			200	1.00	0.04	0.79	0.04	0.81	1.00	0.05	0.81	0.05	0.85
			500	1.00	0.04	0.99	0.04	0.99	1.00	0.05	0.99	0.05	1.00
		L	50	1.00	0.04	0.77	0.05	0.83	1.00	0.04	0.74	0.06	0.84
			100	1.00	0.04	0.98	0.05	0.98	1.00	0.05	0.99	0.05	0.99
			200	1.00	0.05	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
			500	1.00	0.04	1.00	0.04	1.00	1.00	0.05	1.00	0.05	1.00
	IT	S	50	1.00	0.05	0.46	0.07	0.52	1.00	0.04	0.42	0.05	0.50
			100	1.00	0.04	0.80	0.05	0.82	1.00	0.05	0.75	0.05	0.79
			200	1.00	0.04	0.98	0.04	0.98	1.00	0.05	0.97	0.05	0.98
			500	1.00	0.06	1.00	0.06	1.00	1.00	0.05	1.00	0.05	1.00
		L	50	1.00	0.05	0.96	0.07	0.97	1.00	0.04	0.94	0.05	0.97
			100	1.00	0.04	1.00	0.05	1.00	1.00	0.05	1.00	0.05	1.00
			200	1.00	0.04	1.00	0.04	1.00	1.00	0.05	1.00	0.05	1.00
			500	1.00	0.06	1.00	0.06	1.00	1.00	0.05	1.00	0.05	1.00

Note. PN= Percent of Non-invariance; SN=Source of Non-invariance; LD=Factor Loadings; IT=Intercept; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select an item as RI. ASR=Admissible Solution Rate. ML_LR represents the likelihood ratio test using maximum likelihood estimation (MLE). Under MLE, the admissible solution rate was 1.00 across all listed conditions.

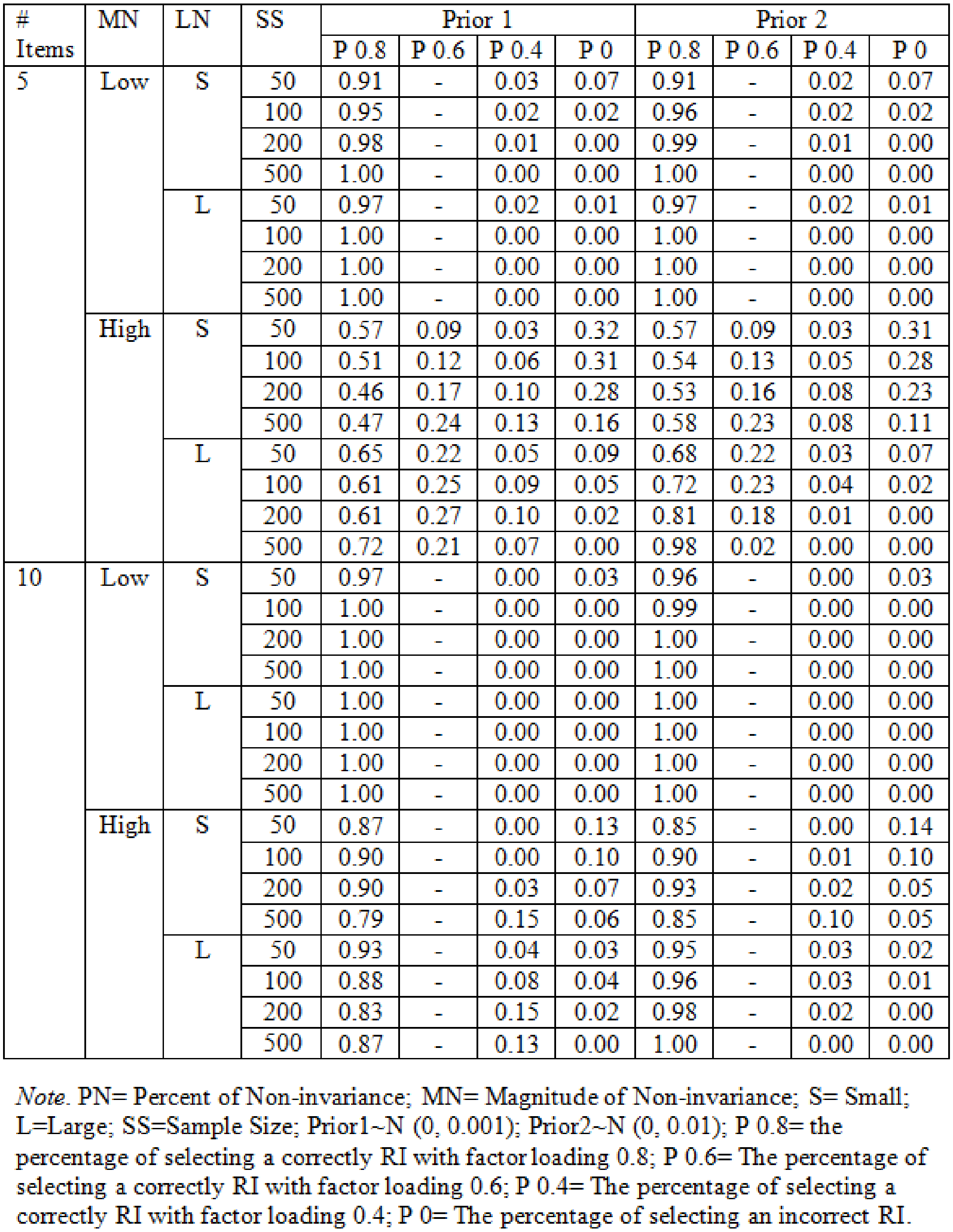
Table 20: Power Rates of Selecting Three Reference Indicators



# OF Item	MN	PN	SS	Power1				Power2			
				AR	BSEM		AOAR Max3	AR	BSEM		AOAR Max3
					Prior1	Prior2			Prior1	Prior2	
5 Items	S	Low	50	0.40	0.66	0.67	0.31	0.60	0.77	0.78	0.54
			100	0.40	0.80	0.82	0.42	0.60	0.87	0.88	0.61
			200	0.40	0.94	0.94	0.66	0.60	0.96	0.96	0.78
			500	0.40	1.00	1.00	0.90	0.60	1.00	1.00	0.93
		High	50	0.10	0.05	0.05	0.02	0.53	0.63	0.64	0.44
			100	0.10	0.04	0.05	0.00	0.53	0.66	0.66	0.43
			200	0.10	0.05	0.06	0.01	0.53	0.68	0.68	0.47
			500	0.10	0.09	0.14	0.01	0.53	0.69	0.71	0.37
10 Items	S	Low	50	0.47	0.71	0.70	0.13	0.67	0.90	0.89	0.63
			100	0.47	0.90	0.89	0.31	0.67	0.97	0.96	0.72
			200	0.47	0.96	0.94	0.45	0.67	0.99	0.98	0.79
			500	0.47	1.00	1.00	0.94	0.67	1.00	1.00	0.98
		High	50	0.17	0.17	0.16	0.01	0.60	0.63	0.62	0.25
			100	0.17	0.15	0.13	0.00	0.60	0.66	0.64	0.19
			200	0.17	0.23	0.19	0.01	0.60	0.72	0.70	0.25
			500	0.17	0.48	0.33	0.17	0.60	0.82	0.76	0.51

Note. PN= Percent of Non-invariance; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; AR= power rates as randomly select three items as RIs. For the BSEM method, Prior1~N (0, 0.001); Prior2~N(0, 0.01); Prior3~N (0, 0.001); Prior2~N(0, 0.01); AOAR Max3 represents the method of using all other items as reference indicators (i.e. constrained baseline approach) to screen out possible invariant items, and then selecting three items with the largest factor loadings as RIs. Power1 was calculated as the percentage of selecting all three RIs correctly; Power2 was calculated as the probability of being truly invariant among the three selected RIs (e.g. 0.7 indicates 70% of the selected RIs are expected to be truly invariant).

Table 21: Power Rates of Selecting RI with Different Factor Loadings



# Items	MN	LN	SS	Prior 1				Prior 2			
				P 0.8	P 0.6	P 0.4	P 0	P 0.8	P 0.6	P 0.4	P 0
5	Low	S	50	0.91	-	0.03	0.07	0.91	-	0.02	0.07
			100	0.95	-	0.02	0.02	0.96	-	0.02	0.02
			200	0.98	-	0.01	0.00	0.99	-	0.01	0.00
			500	1.00	-	0.00	0.00	1.00	-	0.00	0.00
		L	50	0.97	-	0.02	0.01	0.97	-	0.02	0.01
			100	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			200	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			500	1.00	-	0.00	0.00	1.00	-	0.00	0.00
	High	S	50	0.57	0.09	0.03	0.32	0.57	0.09	0.03	0.31
			100	0.51	0.12	0.06	0.31	0.54	0.13	0.05	0.28
			200	0.46	0.17	0.10	0.28	0.53	0.16	0.08	0.23
			500	0.47	0.24	0.13	0.16	0.58	0.23	0.08	0.11
		L	50	0.65	0.22	0.05	0.09	0.68	0.22	0.03	0.07
			100	0.61	0.25	0.09	0.05	0.72	0.23	0.04	0.02
			200	0.61	0.27	0.10	0.02	0.81	0.18	0.01	0.00
			500	0.72	0.21	0.07	0.00	0.98	0.02	0.00	0.00
10	Low	S	50	0.97	-	0.00	0.03	0.96	-	0.00	0.03
			100	1.00	-	0.00	0.00	0.99	-	0.00	0.00
			200	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			500	1.00	-	0.00	0.00	1.00	-	0.00	0.00
		L	50	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			100	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			200	1.00	-	0.00	0.00	1.00	-	0.00	0.00
			500	1.00	-	0.00	0.00	1.00	-	0.00	0.00
	High	S	50	0.87	-	0.00	0.13	0.85	-	0.00	0.14
			100	0.90	-	0.00	0.10	0.90	-	0.01	0.10
			200	0.90	-	0.03	0.07	0.93	-	0.02	0.05
			500	0.79	-	0.15	0.06	0.85	-	0.10	0.05
		L	50	0.93	-	0.04	0.03	0.95	-	0.03	0.02
			100	0.88	-	0.08	0.04	0.96	-	0.03	0.01
			200	0.83	-	0.15	0.02	0.98	-	0.02	0.00
			500	0.87	-	0.13	0.00	1.00	-	0.00	0.00

Note. PN= Percent of Non-invariance; MN= Magnitude of Non-invariance; S= Small; L=Large; SS=Sample Size; Prior1~N (0, 0.001); Prior2~N (0, 0.01); P 0.8= the percentage of selecting a correctly RI with factor loading 0.8; P 0.6= The percentage of selecting a correctly RI with factor loading 0.6; P 0.4= The percentage of selecting a correctly RI with factor loading 0.4; P 0= The percentage of selecting an incorrect RI.

Figure 1: Plots of EDOI for Item 1

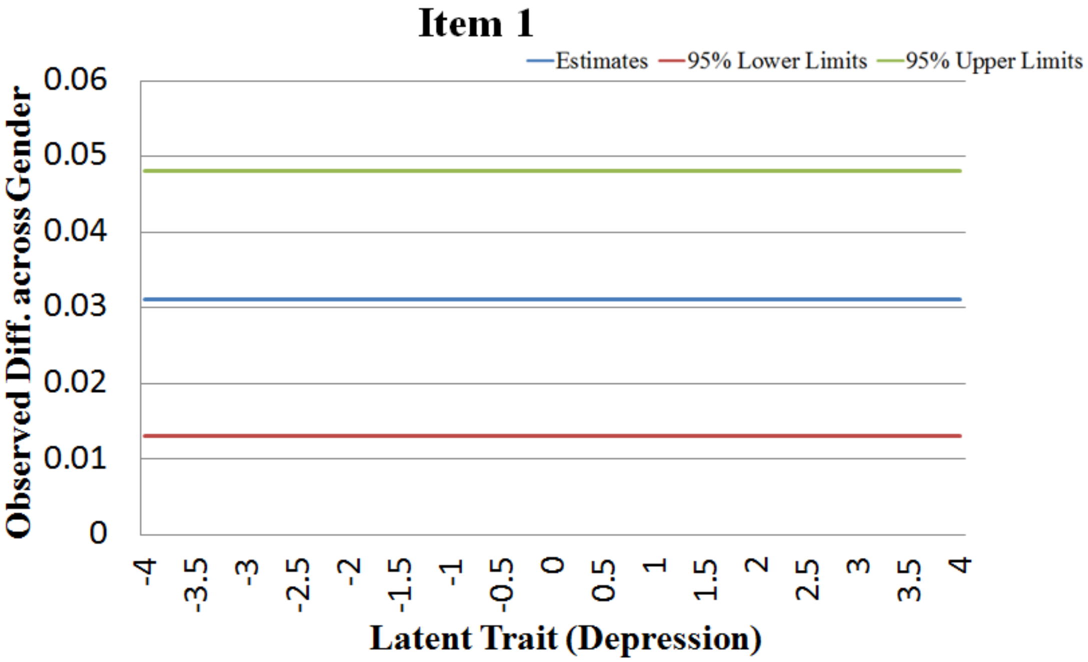


Figure 2: Plots of EDOI for Item 2

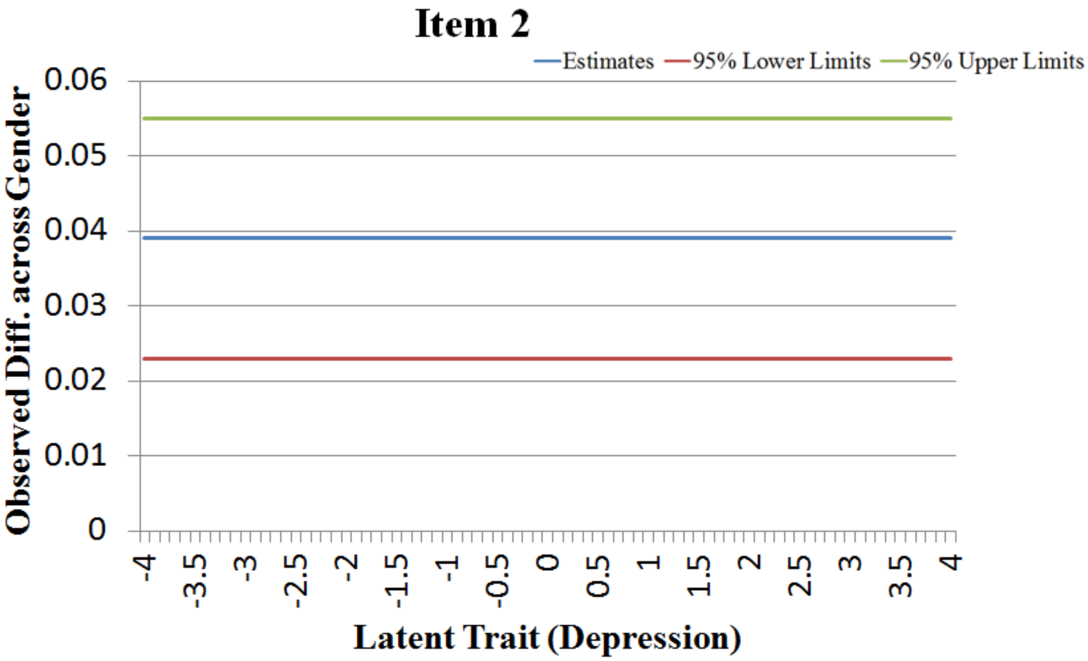


Figure 3: Plots of EDOI for Item 5

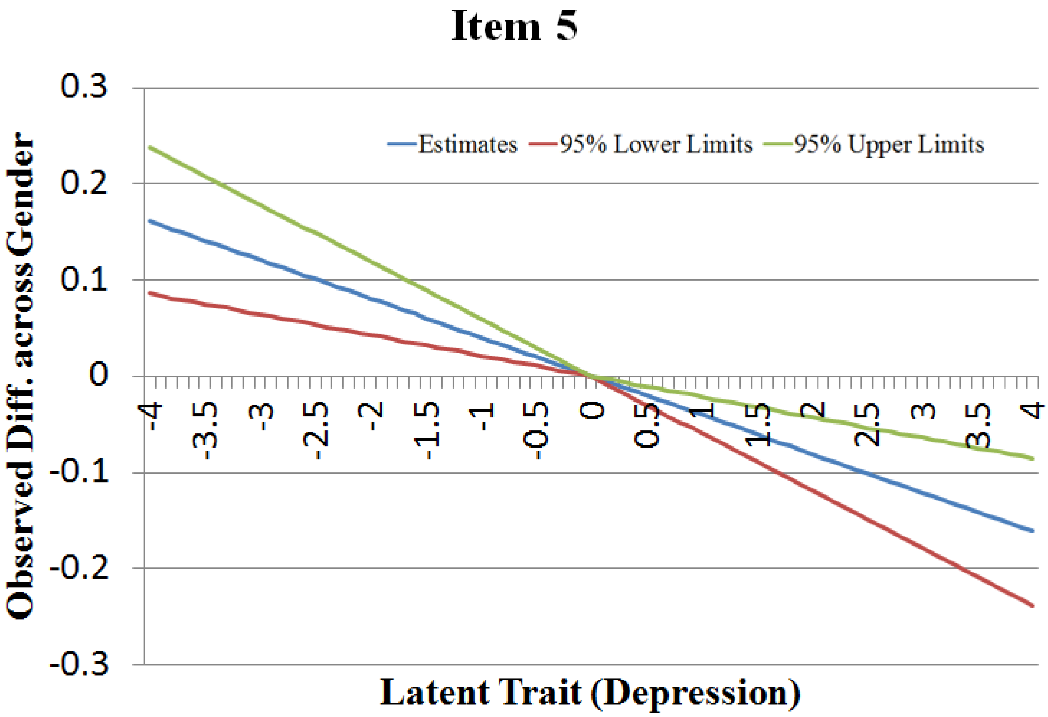


Figure 4: Plots of EDOI for Item 7

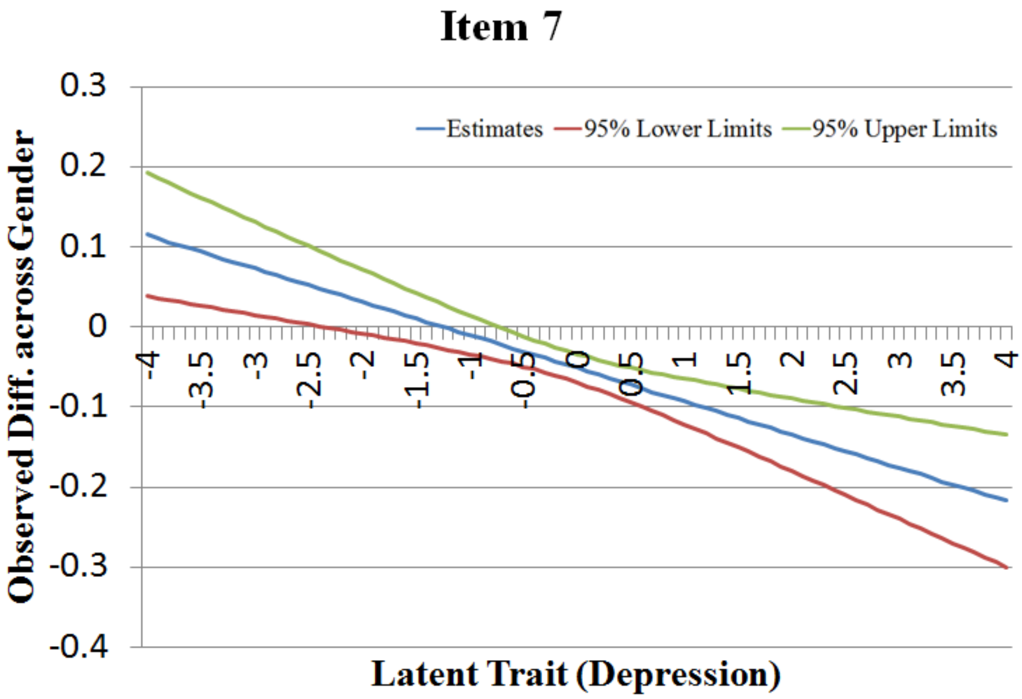


Figure 5: Plots of EDOI for Item 9

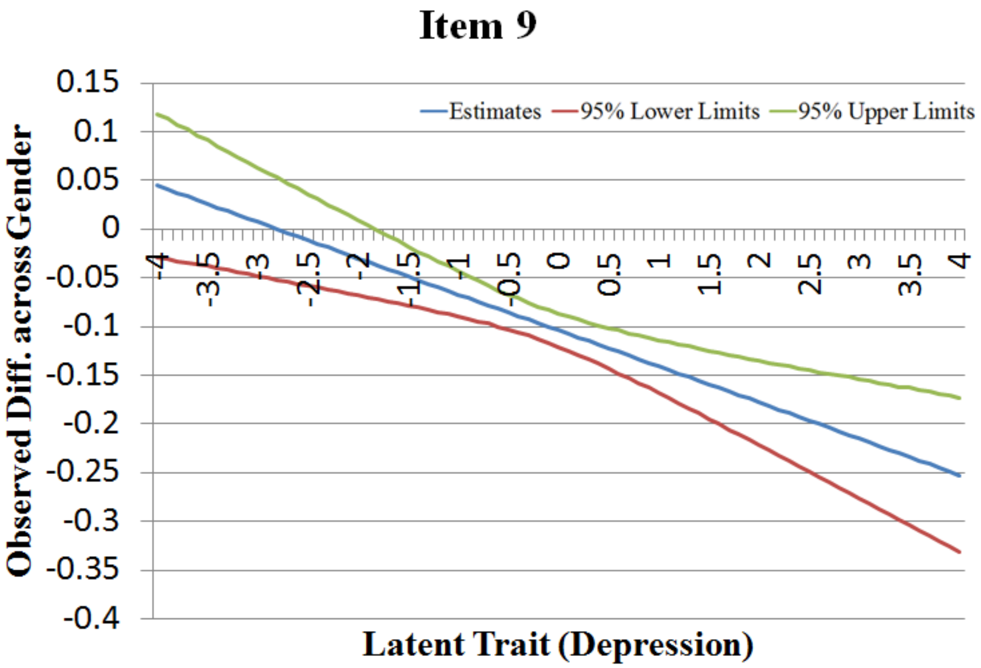


Figure 6: Plots of EDOI for Item 10

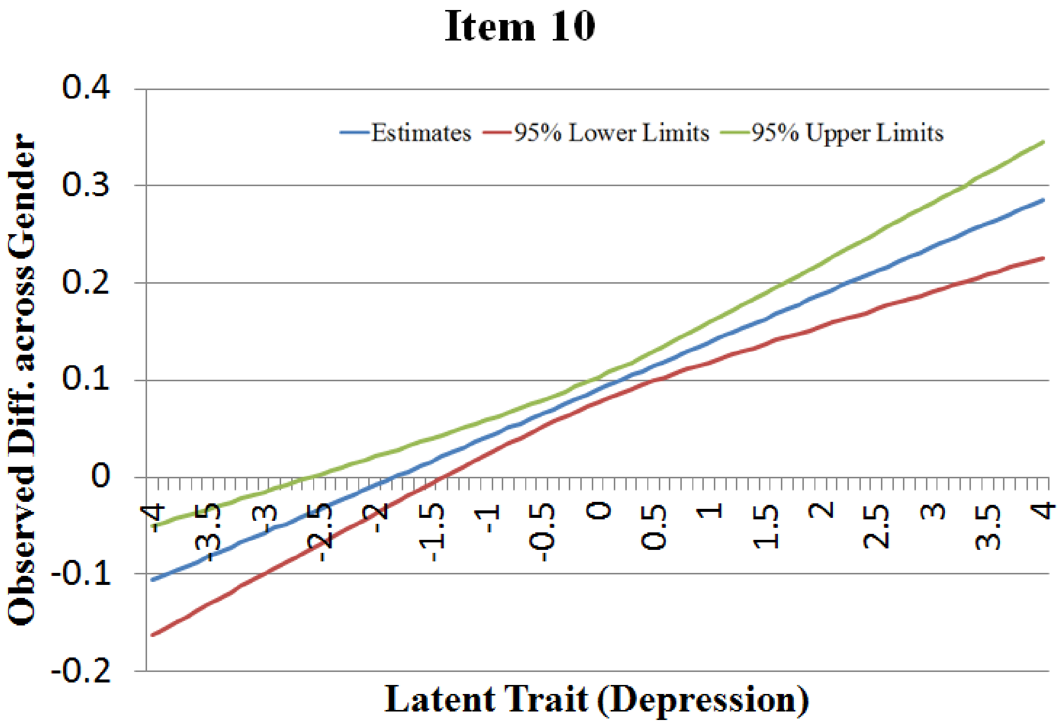


Figure 7: Plots of EDOI for Item 11

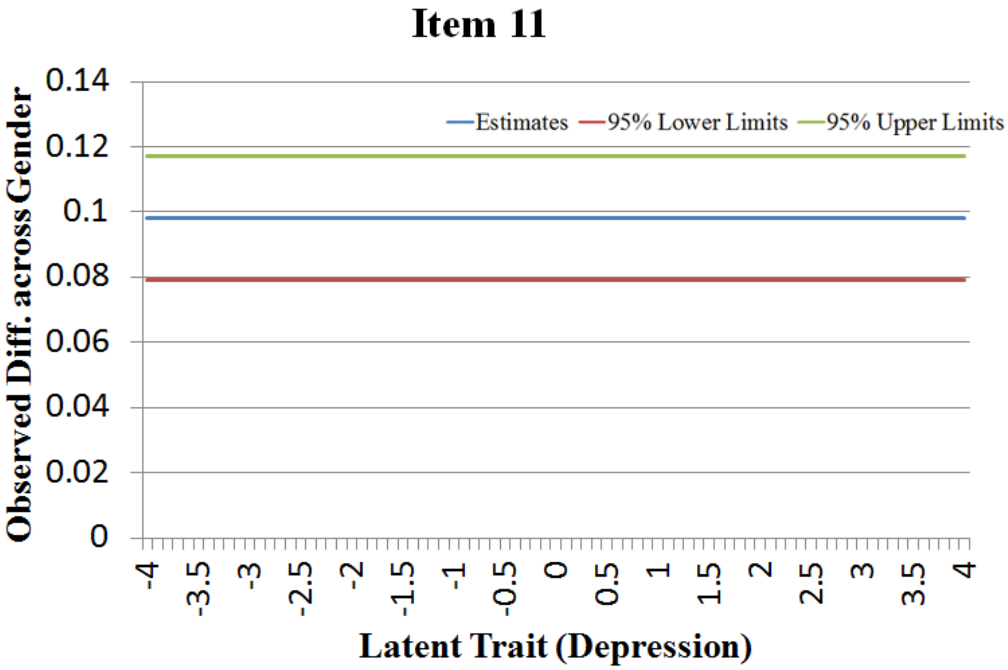


Figure 8: Plots of EDOI for Item 13

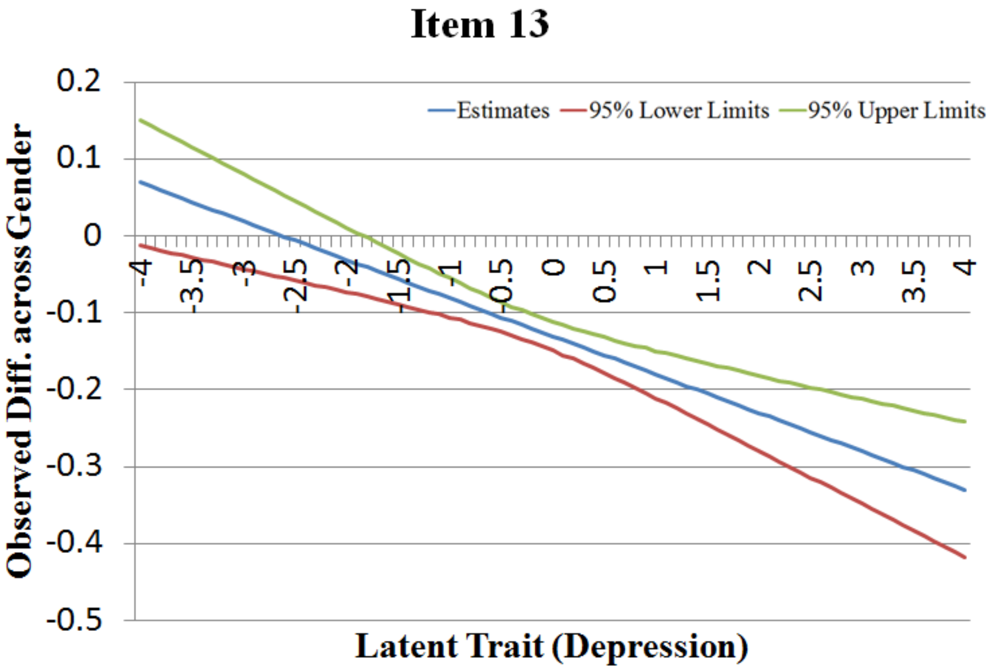


Figure 9: Plots of EDOI for Item 14

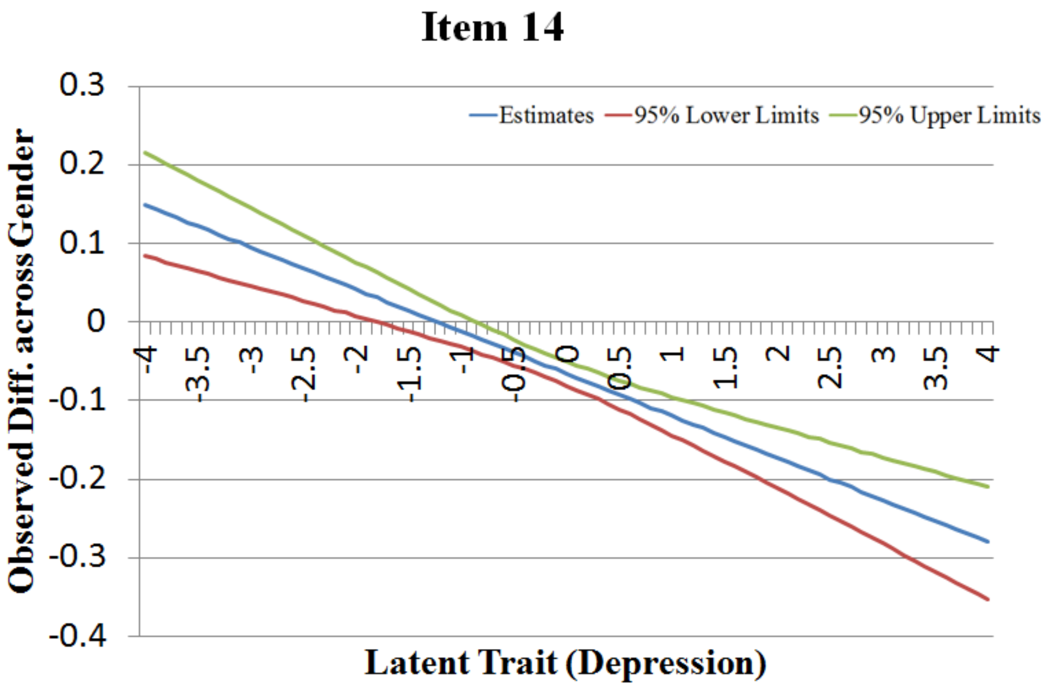


Figure 10: Plots of EDOI for Item 17

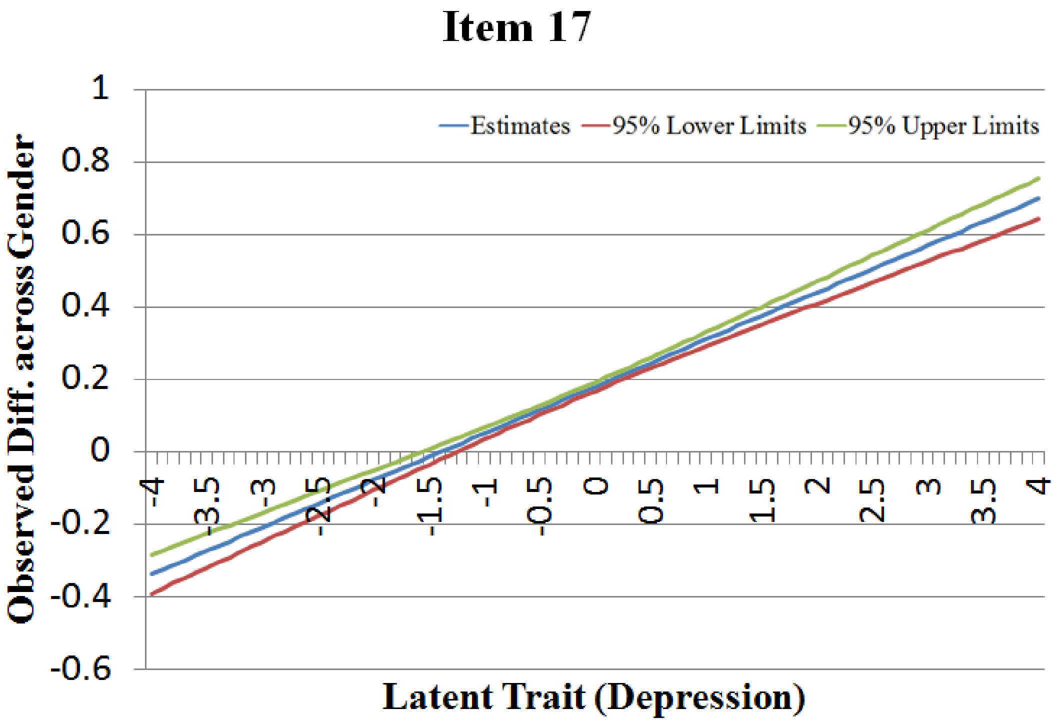


Figure 11: Plots of EDOI for Item 18

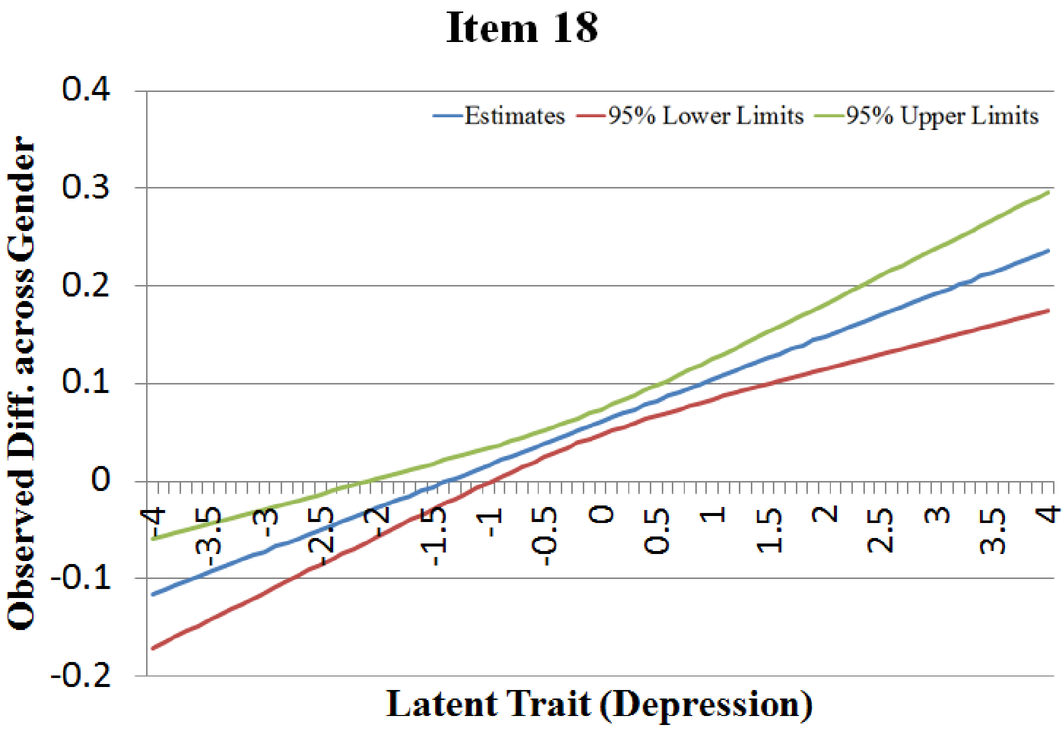


Figure 12: Plots of EDOI for Item 19

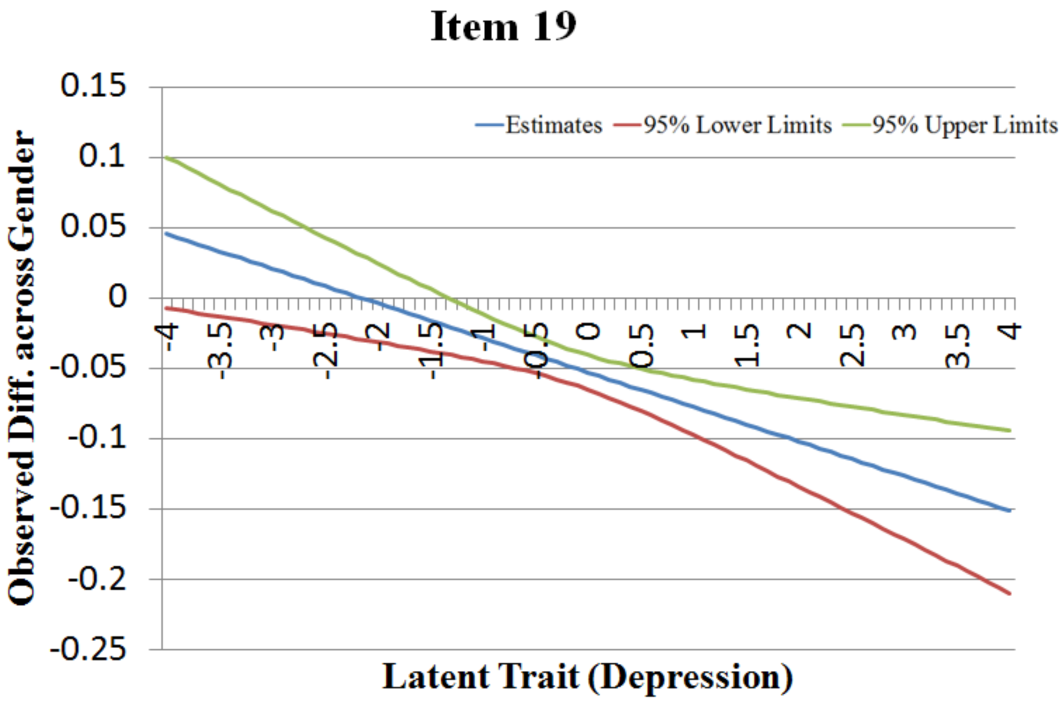


Figure 13: Plots of EDOI for Item 20

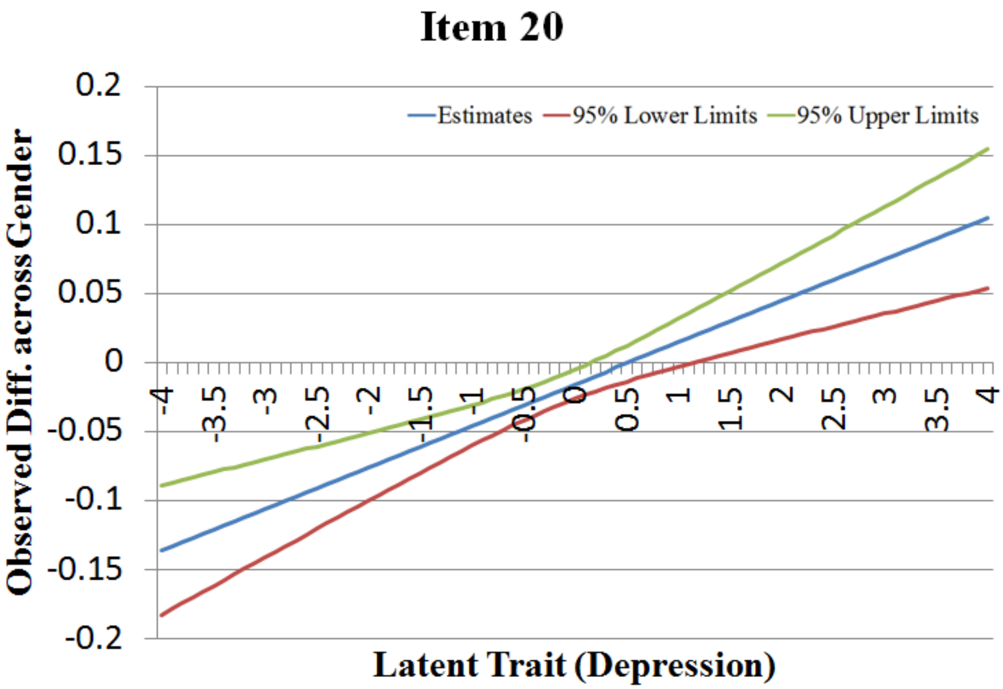


Figure 14: Plots of EDOT

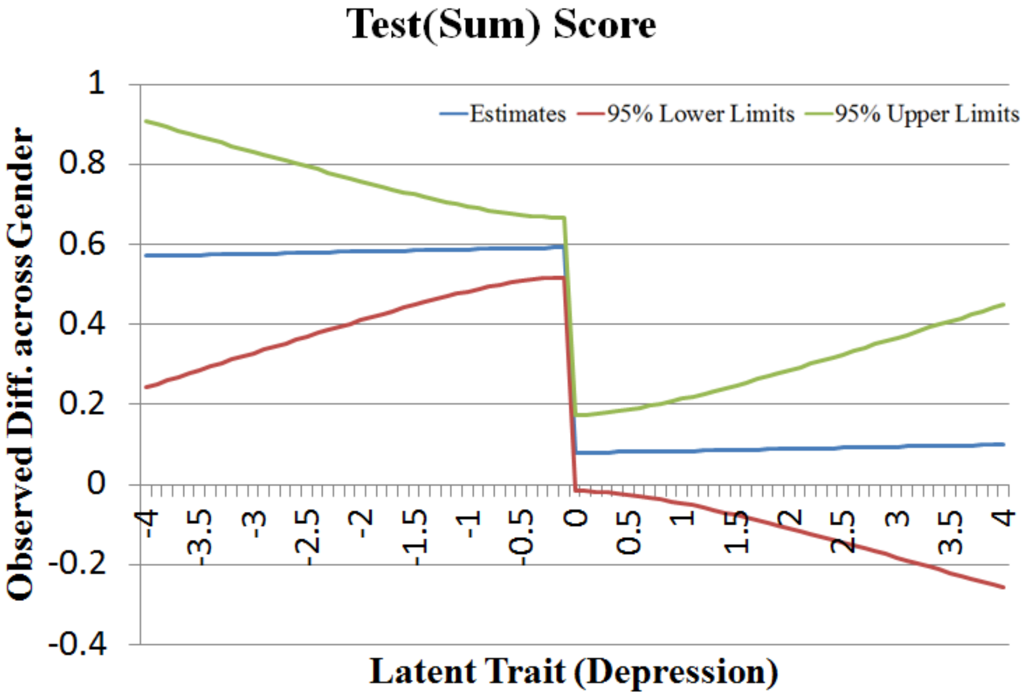
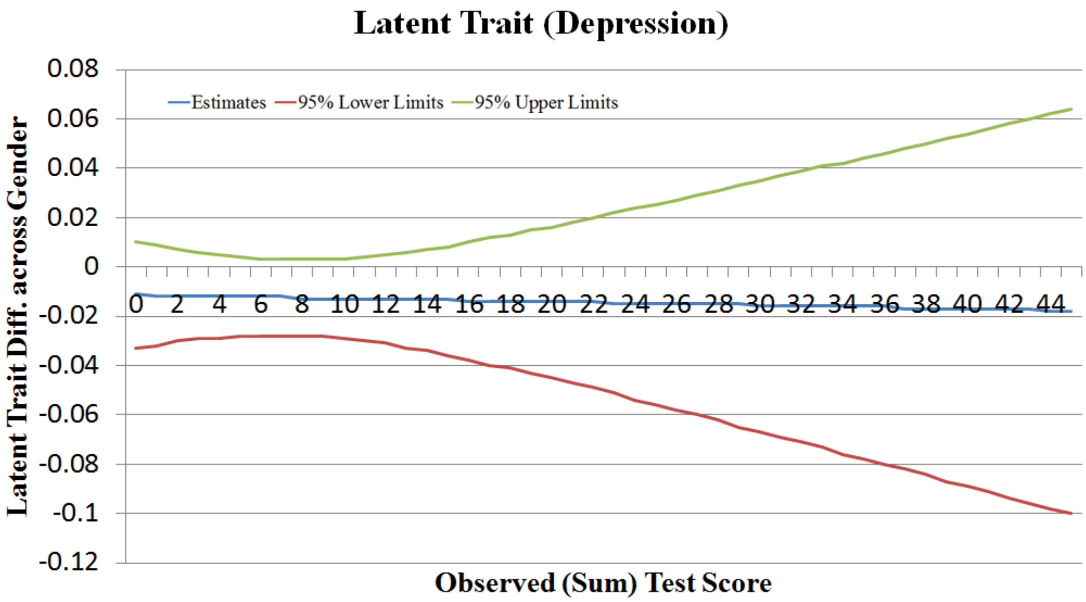


Figure 15: Plots of EDLT



References

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

Bentler, P. M., & Bonett, D. G. (1980). Significance Tests and Goodness of Fit in the Analysis of Covariance Structures. *Psychological Bulletin, 88*(3), 588.

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons. Chicago

Brooks, S. P. (2003). Bayesian computation: a statistical revolution. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 361*(1813), 2681-2697.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the Equivalence of Factor Covariance and Mean Structures - the Issue of Partial Measurement Invariance. *Psychological Bulletin, 105*(3), 456-466.

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

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

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant measurement items. *Applied Behavioral Science Review, 6*(1), 93-110.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*(1), 1-27.

Cohen, J. (1994). The Earth is Round (p<0.05). *American Psychologist, 49,* 997-1003.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662.

Dienes, Z. (2011). Bayesian versus Orthodox Statistics: Which Side are You On?. *Perspectives on Psychological Science, 6*(3), 274-290.

Edwards, M. C., Cheavens, J. S., Heiy, J. E., & Cukrowicz, K. C. (2010). A Reexamination of the Factor Structure of the Center for Epidemiologic Studies Depression Scale: is a One-factor Model Plausible?. *Psychological Assessment, 22*(3), 711.

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(1), 22-36.

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling-a Multidisciplinary Journal, 15*(1), 96-113.

Gagné, P., & Furlow, C. F. (2009). Automating multiple software packages in simulation research for structural equation modeling and hierarchical linear modeling. *Structural equation modeling, 16*(1), 179-185.

Hoogland, J. J., & Boomsma, A. (1998). Robustness Studies in Covariance Structure Modeling An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invarient: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist*.

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The Role of Referent Indicators in Tests of Measurement Invariance. *Structural Equation Modeling-a Multidisciplinary Journal, 16*(4), 642-657.

Jöreskog, K. G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika*, *36*(4), 409-&.

Karkee, T., & Choi, S. (2005). Impact of Eliminating Anchor Items Flagged from Statistical Criteria on Test Score Classifications in Common Item Equating. Paper Presented at the *Annual Meeting of the American Educational Research Association*, Montreal, April 15, 2005

Kim, S. H., & Cohen, A. S. (1995). A comparison of Lord's chi-square, Raju's area measures, and the likelihood ratio test on detection of differential item functioning. *Applied Measurement in Education,8*(4), 291-312.

Kim, E. S., & Yoon, M. (2011). 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., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC: Likelihood Ratio Test With a Critical Value Adjustment. *Educational and Psychological Measurement*, *72*(3), 469-492

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The time has come Bayesian methods for data analysis in the organizational sciences. *Organizational Research Methods, 15*(4), 722-752.

Lee, S. Y., & Song, X. Y. (2004). Bayesian Model Comparison of Nonlinear Structural Equation Models with Missing Continuous and Ordinal Categorical Data. British *Journal of Mathematical and Statistical Psychology, 57*(1), 131-150.

Liao, X., Song, H., & Shi, D. (2015). The Impact of Measurement Non-Equivalence on Second-Order Latent Growth Curve Modeling. *Manuscript in preparation*.

Little, R. J. (2006). Calibrated Bayes: a Bayes/frequentist roadmap. *The American Statistician, 60*(3), 213-223.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A Non-arbitrary Method of Identifying and Scaling Latent Variables in SEM and MACS Models. *Structural Equation Modeling, 13*(1), 59-72.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German Job Satisfaction Survey used in a multinational organization: implications of Schwartz's culture model. *Journal of Applied Psychology,89*(6), 1070-1082.

Martin, A. D. (2005). Bayesian analysis. Retrieved from http://adm.wustl.edu/media/ working/bayes. pdf

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In Maydeu-Olivares A. & McArdle (Eds.), *Contemporary psychometrics* (pp.275-340). Mahwah, NJ:Erlbaum.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490.

MacCallum, R. C., Edwards, M. C., & Cai, L. (2012). Hopes and Cautions in Implementing Bayesian Structural Equation Modeling. *Psychological Methods, 17*(3), 340-345.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G2 (dif) to assess relative model fit in categorical data analysis.*Multivariate Behavioral Research,41(1)*, 55-64.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socio-economic status. *British Journal of Mathematical and Statistical Psychology, 24*, 154-168.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling-a Multidisciplinary Journal, 11*(1), 60-72.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales.*Journal of Applied Psychology,95*(4), 728.

Meade, A. W., & Wright, N. A. (2012). Solving the Measurement Invariance Anchor Item Problem in Item Response Theory. *Journal of Applied Psychology, 97*(5), 1016-1031.

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

Meredith, W. (1993). Measurement Invariance, Factor-Analysis and Factorial Invariance. *Psychometrika*, *58*(4), 525-543.

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

Millsap, R. E. (2005). Four Unresolved Problems in Studies of Factorial Invariance. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary Psychometrics* (pp, 153-170). Mahwah, NJ: Erlbaum.

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. MacCallum (Eds.),*Factor analysis at 100: historical developments and future directions* (pp.131-152). Mahwah, NJ: Erlbaum.

Millsap, R. E. (2011). Statistical Approaches to Measurement Invariance. Routledge.

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.*

Muthén, B., & Asparouhov, T. (2012). Bayesian Structural Equation Modeling: A More Flexible Representation of Substantive Theory. *Psychological Methods, 17*(3), 313-335.

Muthén, B., & Asparouhov, T. (2012b). New Developments in Mplus Version 7. Retrieved from http://www.statmodel.com/download/handouts/MuthenV7Part1.pdf

Muthén, B. O., & Asparouhov, T. (2013). BSEM measurement invariance analysis (MplusWeb Notes No. 17) Retrieved from: http://www.statmodel.com/examples/webnotes/webnote17.pdf.

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

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring Clustering in Confirmatory Factor Analysis: Some Consequences for Model Fit and Standardized Parameter Estimates. *Multivariate Behavioral Research, 49*(6), 518-543.

Radloff, L. S. (1977). The CES-D scale a Self-report Depression Scale for Research in the General Population. *Applied psychological measurement*, *1*(3), 385-401.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory Factor-Analysis and Item Response Theory - 2 Approaches for Exploring Measurement Invariance. *Psychological Bulletin, 114*(3), 552-566.

Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement, 58*(6), 1017-1034.

Rensvold, R. B., & Cheung, G. W. (1999). Testing measurement models for factorial invariance: A Systematic approach (vol 58, pg 1032, 1998). *Educational and Psychological Measurement, 59*(1), 186-186.

Rivas, G. E. L., 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,* 251-265.

Schmitt, N., Golubovich, J., & Leong, F. T. (2010). Impact of measurement invariance on construct correlations, mean differences, and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*.

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

Serang, S., Zhang, Z., Helm, J., Steele, J. S., & Grimm, K. J. (2015). Evaluation of a Bayesian Approach to Estimating Nonlinear Mixed-Effects Mixture Models. *Structural Equation Modeling: A Multidisciplinary,22*(2), 202-215.

Shi, D., Song, H., & Lewis, M. (2016). Impact of Partial Factorial Invariance on Cross-Group Comparison. *Manuscrpt under review*.

Shevlin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: a large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17*(2), 231.

Song, X.-Y., Lee, S.-Y., & Wiley InterScience (Online service). (2012). *Basic and advanced Bayesian structural equation modeling with applications in the medical and behavioral sciences Wiley series in probability and statistics* (pp. 1 online resource.). Retrieved from http://libraries.ou.edu/access.aspx?url=http://onlinelibrary.wiley.com/book/10.1002/ 9781118358887

Song, H., & Ferrer, E. (2012). Bayesian Estimation of Random Coefficient Dynamic Factor Models. *Multivariate Behavioral Research,47*(1), 26-60.

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.

Steinmetz, H. (2013). Analyzing observed composite differences across groups. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 9*(1), 1-12.

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

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association, 82*(398), 528-540.

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.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. New York: Guilford.

Widaman, K. F. (1993). Common Factor Analysis Versus Principal Component Analysis: Differential Bias in Representing Model Parameters? *Multivariate Behavioral Research, 28*(3), 263-311.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance Within Longitudinal Structural Equation Models: Measuring the Same Construct Across Time. *Child Development Perspectives, 4*(1), 10-18.

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

Woods, C. M., & Grimm, K. J. (2011). Testing for Nonuniform Differential Item Functioning With Multiple Indicator Multiple Cause Models. Applied Psychological Measurement, 35(5), 339-361.

Xie, Y., & Hu, J. (2014). An introduction to the China family panel studies (CFPS). *Chinese Sociological Review, 47*(1), 3-29.

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.

Appendices

Appendix A: Footnotes

1. Alternatively, one can begin the tests by fitting a model with all parameters constrained to be equal, and then progressively relaxing certain equality constraints. More discussions between these two approaches are discussed in the later section, and can be found in Stark, Chernyshenko and Drasgow (2006) and Kim and Yoon (2011).
2. See Little, Slegers and Card (2006) for a detailed discussion on the issue of identifying and scaling latent variables in multiple-group models.
3. In this study, we focus on testing measurement invariance across two groups. The proposed BSEM method can be naturally extended to testing for longitudinal factorial invariance.
4. Equations 8-10 can be derived from Equation 2.
5. The other methods, such as those we reviewed in the manuscripts, are not quite comparable to the BSEM method regarding to the selection of RI. In general, those methods attempt to avoid using a specific RI for invariance test. Therefore, selecting RIs is not their goal to achieve. For example, Cheung and Lau (2012)'s method detects non-invariance without selecting specific RIs. Therefore, no additional simulations can be done to compare the BSEM with other methods like Cheung and Lau's.
6. The standard deviation for the female group is set to be one.

7. The thresholds of depression level that suggesting whether females have higher observed scores (than males) are -1.2 (item 7), -2.8(item 9), -2.6 (item 13), -1.2 (item 14), and -2.1 (item 19).
8. The thresholds of depression level that suggesting whether females have lower observed scores (than males) are -1.8 (item10), and -1.4 (item 18).
9. For the group of females, the latent trait (depression) has mean 0 and variance 1, for males, the latent trait (depression) has mean -0.281 and variance 0.752.

Appendix B: Mplus Syntax for selecting RI

```
TITLE: This is an example of selecting RI for a fifteen-
item scale
DATA: FILE IS data.dat;
VARIABLE: NAMES ARE x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14
x17 x18 x19 x20 sex;
CLASSES=c(2);
KNOWNCLASS=c(sex=1 2);

ANALYSIS:
type=mixture;
estimator=bayes;
proc=2;
thin=10;
biterations =100000(50000);

MODEL:
    %OVERALL%
f BY x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14 x17 x18 x19 x20;
[x1*];
[x2*];
[x3*];
[x5*];
[x6*];
[x7*];
[x9*];
[x10*];
[x11*];
[x13*];
[x14*];
[x17*];
[x18*];
[x19*];
[x20*];
x1*;
x2*;
x3*;
x5*;
x6*;
x7*;
x9*;
x10*;
x11*;
x13*;
x14*;
x17*;
```

```
x18*;
x19*;
x20*;

%c#1%
f BY   x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D1)
       x6*(E1)
       x7*(F1)
       x9*(G1)
       x10*(H1)
       x11*(I1)
       x13*(J1)
       x14*(K1)
       x17*(L1)
       x18*(M1)
       x19*(N1)
       x20*(O1);
      [x1*](P1)
      [x2*](Q1)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U1)
      [x9*](V1)
      [x10*](W1)
      [x11*](X1)
      [x13*](Y1)
      [x14*](Z1)
      [x17*](AA1)
      [x18*](BB1)
      [x19*](CC1)
      [x20*](DD1);
      f@1;
      [f@0];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;
%c#2%
f BY   x1*(A2)
       x2*(B2)
       x3*(C2)
       x5*(D2)
       x6*(E2)
       x7*(F2)
       x9*(G2)
       x10*(H2)
       x11*(I2)
       x13*(J2)
       x14*(K2)
       x17*(L2)
       x18*(M2)
       x19*(N2)
       x20*(O2);
      [x1*](P2)
      [x2*](Q2)
      [x3*](R2)
      [x5*](S2)
      [x6*](T2)
      [x7*](U2)
      [x9*](V2)
      [x10*](W2)
      [x11*](X2)
      [x13*](Y2)
      [x14*](Z2)
      [x17*](AA2)
      [x18*](BB2)
      [x19*](CC2)
      [x20*](DD2);
      f*;
      [f*];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

MODEL CONSTRAINT:
    new(Dif1-Dif30*0);
     Dif1=A1-A2;
     Dif2=B1-B2;
     Dif3=C1-C2;
     Dif4=D1-D2;
     Dif5=E1-E2;
     Dif6=F1-F2;
     Dif7=G1-G2;
     Dif8=H1-H2;
     Dif9=I1-I2;
     Dif10=J1-J2;
     Dif11=k1-k2;
     Dif12=L1-L2;
     Dif13=M1-M2;
     Dif14=N1-N2;
     Dif15=O1-O2;
     Dif16=P1-P2;
     Dif17=Q1-Q2;
     Dif18=R1-R2;
     Dif19=S1-S2;
     Dif20=T1-T2;
     Dif21=U1-U2;
     Dif22=V1-V2;
     Dif23=W1-W2;
     Dif24=X1-X2;
     Dif25=Y1-Y2;
     Dif26=Z1-Z2;
     Dif27=AA1-AA2;
     Dif28=BB1-BB2;
     Dif29=CC1-CC2;
     Dif30=DD1-DD2;
MODEL PRIOR:
    DIFF(A1,A2)~N(0,0.01);
    DIFF(B1,B2)~N(0,0.01);
    DIFF(C1,C2)~N(0,0.01);
    DIFF(D1,D2)~N(0,0.01);
    DIFF(E1,E2)~N(0,0.01);
    DIFF(F1,F2)~N(0,0.01);
    DIFF(G1,G2)~N(0,0.01);
DIFF(H1,H2)~N(0,0.01);
DIFF(I1,I2)~N(0,0.01);
DIFF(J1,J2)~N(0,0.01);
DIFF(K1,K2)~N(0,0.01);
DIFF(L1,L2)~N(0,0.01);
DIFF(M1,M2)~N(0,0.01);
DIFF(N1,N2)~N(0,0.01);
DIFF(O1,O2)~N(0,0.01);
DIFF(P1,P2)~N(0,0.01);
DIFF(Q1,Q2)~N(0,0.01);
DIFF(R1,R2)~N(0,0.01);
DIFF(S1,S2)~N(0,0.01);
DIFF(T1,T2)~N(0,0.01);
DIFF(U1,U2)~N(0,0.01);
DIFF(V1,V2)~N(0,0.01);
DIFF(W1,W2)~N(0,0.01);
DIFF(X1,X2)~N(0,0.01);
DIFF(Y1,Y2)~N(0,0.01);
DIFF(Z1,Z2)~N(0,0.01);
DIFF(AA1,AA2)~N(0,0.01);
DIFF(BB1,BB2)~N(0,0.01);
DIFF(CC1,CC2)~N(0,0.01);
DIFF(DD1,DD2)~N(0,0.01);
```

Appendix C: Mplus Syntax for locating non-invariant parameters

```
TITLE: This is an example of locating non-invariant
parameters(using Item 6 as RI)

DATA: FILE IS data.dat;
VARIABLE: NAMES ARE x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14
x17 x18 x19 x20 sex;
CLASSES=c(2);
KNOWNCLASS=c(sex=1 2);

ANALYSIS:
type=mixture;
estimator=bayes;
proc=2;
thin=10;
biterations =100000(50000);

MODEL:
%OVERALL%
f BY x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14 x17 x18 x19 x20;
[x1*];
[x2*];
[x3*];
[x5*];
[x6*];
[x7*];
[x9*];
[x10*];
[x11*];
[x13*];
[x14*];
[x17*];
[x18*];
[x19*];
[x20*];
x1*;
x2*;
x3*;
x5*;
x6*;
x7*;
x9*;
x10*;
x11*;
x13*;
x14*;
```

```
x17*;
x18*;
x19*;
x20*;


    %c#1%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D1)
       x6*(E1)
       x7*(F1)
       x9*(G1)
       x10*(H1)
       x11*(I1)
       x13*(J1)
       x14*(K1)
       x17*(L1)
       x18*(M1)
       x19*(N1)
       x20*(O1);
      [x1*](P1)
      [x2*](Q1)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U1)
      [x9*](V1)
      [x10*](W1)
      [x11*](X1)
      [x13*](Y1)
      [x14*](Z1)
      [x17*](AA1)
      [x18*](BB1)
      [x19*](CC1)
      [x20*](DD1);
      f@1;
      [f@0];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
    x11*;
    x13*;
    x14*;
    x17*;
    x18*;
    x19*;
    x20*;

 %c#2%
f BY x1*(A2)
     x2*(B2)
     x3*(C2)
     x5*(D2)
     x6*(E1)
     x7*(F2)
     x9*(G2)
     x10*(H2)
     x11*(I2)
     x13*(J2)
     x14*(K2)
     x17*(L2)
     x18*(M2)
     x19*(N2)
     x20*(O2);
    [x1*](P2)
    [x2*](Q2)
    [x3*](R2)
    [x5*](S2)
    [x6*](T1)
    [x7*](U2)
    [x9*](V2)
    [x10*](W2)
    [x11*](X2)
    [x13*](Y2)
    [x14*](Z2)
    [x17*](AA2)
    [x18*](BB2)
    [x19*](CC2)
    [x20*](DD2);
    f*;
    [f*];
    x1*;
    x2*;
    x3*;
    x5*;
    x6*;
    x7*;
```

```
     x9*;
     x10*;
     x11*;
     x13*;
     x14*;
     x17*;
     x18*;
     x19*;
     x20*;

MODEL CONSTRAINT:
  new(Dif1-Dif28*0);
   Dif1=A1-A2;
   Dif2=B1-B2;
   Dif3=C1-C2;
   Dif4=D1-D2;
   Dif5=F1-F2;
   Dif6=G1-G2;
   Dif7=H1-H2;
   Dif8=I1-I2;
   Dif9=J1-J2;
   Dif10=k1-k2;
   Dif11=L1-L2;
   Dif12=M1-M2;
   Dif13=N1-N2;
   Dif14=O1-O2;
   Dif15=P1-P2;
   Dif16=Q1-Q2;
   Dif17=R1-R2;
   Dif18=S1-S2;
   Dif19=U1-U2;
   Dif20=V1-V2;
   Dif21=W1-W2;
   Dif22=X1-X2;
   Dif23=Y1-Y2;
   Dif24=Z1-Z2;
   Dif25=AA1-AA2;
   Dif26=BB1-BB2;
   Dif27=CC1-CC2;
   Dif28=DD1-DD2;
```

Appendix D: Mplus Syntax for Estimating EDOI (Item17, ξ=-4.00)

```
DATA: FILE IS data.dat;
VARIABLE: NAMES ARE x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14
x17 x18 x19 x20 sex;
CLASSES=c(2);
KNOWNCLASS=c(sex=1 2);

ANALYSIS:
type=mixture;
estimator=bayes;
proc=2;
thin=10;
biterations =100000(50000);

MODEL:
%OVERALL%
f BY x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14 x17 x18 x19 x20;
[x1*];
[x2*];
[x3*];
[x5*];
[x6*];
[x7*];
[x9*];
[x10*];
[x11*];
[x13*];
[x14*];
[x17*];
[x18*];
[x19*];
[x20*];
x1*;
x2*;
x3*;
x5*;
x6*;
x7*;
x9*;
x10*;
x11*;
x13*;
x14*;
x17*;
x18*;
x19*;
```

```
x20*;

    %c#1%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D1)
       x6*(E1)
       x7*(F1)
       x9*(G1)
       x10*(H1)
       x11*(I1)
       x13*(J1)
       x14*(K1)
       x17*(L1)
       x18*(M1)
       x19*(N1)
       x20*(O1);
      [x1*](P1)
      [x2*](Q1)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U1)
      [x9*](V1)
      [x10*](W1)
      [x11*](X1)
      [x13*](Y1)
      [x14*](Z1)
      [x17*](AA1)
      [x18*](BB1)
      [x19*](CC1)
      [x20*](DD1);
      f@1;
      [f@0];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

   %c#2%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D2)
       x6*(E1)
       x7*(F2)
       x9*(G2)
       x10*(H2)
       x11*(I1)
       x13*(J2)
       x14*(K2)
       x17*(L2)
       x18*(M2)
       x19*(N2)
       x20*(O2);
      [x1*](P2)
      [x2*](Q2)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U2)
      [x9*](V2)
      [x10*](W2)
      [x11*](X2)
      [x13*](Y2)
      [x14*](Z2)
      [x17*](AA2)
      [x18*](BB2)
      [x19*](CC2)
      [x20*](DD2);
      f*;
      [f*];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
```

```
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

  MODEL CONSTRAINT:
    new(Dif1*0);
Dif1=-4*(L1-L2)+(AA1-AA2);
```

Appendix E: Mplus Syntax for Estimating EDOT (ξ =-4.00)

```
DATA: FILE IS data.dat;
VARIABLE: NAMES ARE x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14
x17 x18 x19 x20 sex;
CLASSES=c(2);
KNOWNCLASS=c(sex=1 2);

ANALYSIS:
type=mixture;
estimator=bayes;
proc=2;
thin=10;
biterations =100000(50000);

MODEL:
%OVERALL%
f BY x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14 x17 x18 x19 x20;
[x1*];
[x2*];
[x3*];
[x5*];
[x6*];
[x7*];
[x9*];
[x10*];
[x11*];
[x13*];
[x14*];
[x17*];
[x18*];
[x19*];
[x20*];
x1*;
x2*;
x3*;
x5*;
x6*;
x7*;
x9*;
x10*;
x11*;
x13*;
x14*;
x17*;
x18*;
x19*;
```

```
x20*;

    %c#1%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D1)
       x6*(E1)
       x7*(F1)
       x9*(G1)
       x10*(H1)
       x11*(I1)
       x13*(J1)
       x14*(K1)
       x17*(L1)
       x18*(M1)
       x19*(N1)
       x20*(O1);
      [x1*](P1)
      [x2*](Q1)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U1)
      [x9*](V1)
      [x10*](W1)
      [x11*](X1)
      [x13*](Y1)
      [x14*](Z1)
      [x17*](AA1)
      [x18*](BB1)
      [x19*](CC1)
      [x20*](DD1);
      f@1;
      [f@0];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

   %c#2%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D2)
       x6*(E1)
       x7*(F2)
       x9*(G2)
       x10*(H2)
       x11*(I1)
       x13*(J2)
       x14*(K2)
       x17*(L2)
       x18*(M2)
       x19*(N2)
       x20*(O2);
      [x1*](P2)
      [x2*](Q2)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U2)
      [x9*](V2)
      [x10*](W2)
      [x11*](X2)
      [x13*](Y2)
      [x14*](Z2)
      [x17*](AA2)
      [x18*](BB2)
      [x19*](CC2)
      [x20*](DD2);
      f*;
      [f*];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
```

```
   x14*;
   x17*;
   x18*;
   x19*;
   x20*;

 MODEL CONSTRAINT:
  new(Dif1*0);
Dif1=P1-P2+Q1-Q2-4*(D1-D2+F1-F2+G1-G2+H1-H2+J1-J2+K1-
K2+L1-L2+M1-M2+N1-N2+O1-O2)+U1-U2+V1-V2+W1-W2+X1-X2+Y1-
Y2+Z1-Z2+AA1-AA2+BB1-BB2+CC1-CC2+DD1-DD2;
```

Appendix F: Mplus Syntax for Estimating EDLT (Sum Test Score=7)

```
DATA: FILE IS data.dat;
VARIABLE: NAMES ARE x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14
x17 x18 x19 x20 sex;
CLASSES=c(2);
KNOWNCLASS=c(sex=1 2);

ANALYSIS:
type=mixture;
estimator=bayes;
proc=2;
thin=10;
biterations =100000(50000);

MODEL:
%OVERALL%
f BY x1 x2 x3 x5 x6 x7 x9 x10 x11 x13 x14 x17 x18 x19 x20;
[x1*];
[x2*];
[x3*];
[x5*];
[x6*];
[x7*];
[x9*];
[x10*];
[x11*];
[x13*];
[x14*];
[x17*];
[x18*];
[x19*];
[x20*];
x1*;
x2*;
x3*;
x5*;
x6*;
x7*;
x9*;
x10*;
x11*;
x13*;
x14*;
x17*;
x18*;
x19*;
```

```
x20*;

    %c#1%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D1)
       x6*(E1)
       x7*(F1)
       x9*(G1)
       x10*(H1)
       x11*(I1)
       x13*(J1)
       x14*(K1)
       x17*(L1)
       x18*(M1)
       x19*(N1)
       x20*(O1);
      [x1*](P1)
      [x2*](Q1)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U1)
      [x9*](V1)
      [x10*](W1)
      [x11*](X1)
      [x13*](Y1)
      [x14*](Z1)
      [x17*](AA1)
      [x18*](BB1)
      [x19*](CC1)
      [x20*](DD1);
      f@1;
      [f@0];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

   %c#2%
  f BY x1*(A1)
       x2*(B1)
       x3*(C1)
       x5*(D2)
       x6*(E1)
       x7*(F2)
       x9*(G2)
       x10*(H2)
       x11*(I1)
       x13*(J2)
       x14*(K2)
       x17*(L2)
       x18*(M2)
       x19*(N2)
       x20*(O2);
      [x1*](P2)
      [x2*](Q2)
      [x3*](R1)
      [x5*](S1)
      [x6*](T1)
      [x7*](U2)
      [x9*](V2)
      [x10*](W2)
      [x11*](X2)
      [x13*](Y2)
      [x14*](Z2)
      [x17*](AA2)
      [x18*](BB2)
      [x19*](CC2)
      [x20*](DD2);
      f*;
      [f*];
      x1*;
      x2*;
      x3*;
      x5*;
      x6*;
      x7*;
      x9*;
      x10*;
      x11*;
      x13*;
```

```
      x14*;
      x17*;
      x18*;
      x19*;
      x20*;

  MODEL CONSTRAINT:
    new(Dif1*0);
Dif1=(7-P1-Q1-R1-S1-T1-U1-V1-W1-X1-Y1-Z1-AA1-BB1-CC1-
DD1)/(A1+B1+C1+D1+E1+F1+G1+H1+I1+J1+K1+L1+M1+N1+O1)-(7-P2-
Q2-R1-S1-T1-U2-V2-W2-X2-Y2-Z2-AA2-BB2-CC2-
DD2)/(A1+B1+C1+D2+E1+F2+G2+H2+I1+J2+K2+L2+M2+N2+O2);
```