Evaluating Factorial Invariance: An Interval Estimation Approach Using Bayesian Structural Equation Modeling (BSEM)

Dexin Shi
University of South Carolina
Hairong Song
University of Oklahoma
Christine DiStefano
Alberto Maydeu-Olivares
Heather L. McDaniel
University of South Carolina
Zhehan Jiang
University of Alabama

Correspondence concerning this article should be addressed to Dexin Shi, Dept. of Psychology, University of South Carolina. Barnwell College. 1512 Pendleton St. Columbia, SC 29208. E-mail: shid@mailbox.sc.edu. OR Zhehang Jiang, University of Alabama, 309E Gorgas Library, Tuscaloosa, AL 35487. E-mail: zjiang17@ua.edu. This work was supported in part by the National Science Foundation through Grant No. SES- 1659936 to A. Maydeu-Olivares

Abstract

In this study, we introduce an interval estimation approach based on Bayesian structural equation modeling (BSEM) to evaluate factorial invariance. For each tested parameter, the size of non-invariance with an uncertainty interval (i.e. highest density interval, HDI) is assessed via Bayesian parameter estimation. By comparing the most credible values (i.e. 95% HDI) with a region of practical equivalence (ROPE), the Bayesian approach allows researchers to 1) support the null hypothesis of practical invariance, and 2) examine the practical importance of the non-invariant parameter. Compared to the traditional likelihood ratio test (LRT), simulation results suggested that the proposed Bayesian approach could offer additional insight into evaluating factorial invariance, thus, leading to more informative conclusions. We provide an empirical example to demonstrate the procedures necessary to implement the proposed method in applied research. The importance of and influences on the choice of an appropriate ROPE are discussed.

Keywords: Bayesian SEM, Parameter Estimation, Factorial Invariance; Highest Density Interval (HDI), Region of Practical Equivalence (ROPE)

Evaluating Factorial Invariance: An Interval Estimation Approach using Bayesian Structural Equation Modeling (BSEM):

Measurement invariance is concerned with whether relationships between latent constructs and corresponding observed variables are the same across different groups (e.g., based on nationality, culture, gender, time occasions; Millsap, 2011). Without establishing measurement invariance, observed differences across groups may simply reflect the differences related to the scale under use rather than actual group differences in the constructs that researchers desire to measure. Therefore, in many disciplines, measurement invariance has been increasingly recognized as a prerequisite for conducting cross-group comparisons.

In a structural equation modeling (SEM) framework, multiple-group confirmatory factor analysis (MG-CFA; Jöreskog, 1971; McGaw & Jöreskog, 1971) has been widely used to test measurement invariance by assessing the equivalence of factor models across groups, or factorial invariance (Meredith, 1993). A standard multiple-group CFA model allows each parameter in the factor model to be estimated freely for each group. The model can be expressed as

$$\mathbf{y}^{(j)} = \boldsymbol{\tau}^{(j)} + \Lambda^{(j)}\boldsymbol{\xi}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \quad (1)$$

where *j* represents group membership for the vector of observed variables **y**, implying that all parameters in the model can differ across the *j* groups, $\boldsymbol{\tau}$ represents the intercept vector, $\boldsymbol{\Lambda}$ denotes the factor loading matrix, $\boldsymbol{\xi}$ is the latent score vector, and $\boldsymbol{\varepsilon}$ represents the unique factor vector.

Based on Equation 1, equivalence tests can be conducted on all factor model parameters, including the factor loadings ($\boldsymbol{\Lambda}$), intercepts ($\boldsymbol{\tau}$) and variances of the unique factors ($\boldsymbol{\Theta}$). Researchers have proposed different forms, or levels of factorial invariance (e.g. Byrne, Shavelson, & Muthén, 1989; Horn & McArdle, 1992; Meredith, 1993; Millsap, 2011; Steenkamp

& Baumgartner, 1998; Vandenberg & Lance, 2000). Configural invariance is the weakest form of invariance and is met when the same factor structure (i.e., same number of factors and same salient factor pattern) is found across groups. Three stricter levels of factorial invariance include weak invariance, strong invariance, and strict invariance. These levels of invariance impose increasing model constraints across groups. Weak invariance is achieved when there are equal factor loadings across groups while strong invariance is present if there are both equal factor loadings and equality of intercepts. Strict factorial invariance requires equal factor loadings, intercepts and uniquenesses across groups; however, this is a rather restrictive condition and is not often tested. In addition, achieving strict invariance is not necessary when conducting some major cross-group comparisons (e.g. comparing mean structures, Meredith, 1993). Therefore, in the current study, we only focus on evaluating strong invariance.

The equivalence of model parameters across groups can be tested by comparing the fit between two nested models, one with the equality constraints imposed (M_1) and the other without (M_0). For example, in a test of weak invariance, researchers first fit a baseline model (M_0) where all parameters are freely estimated (except for those constrained for model identification[1]). In addition, a more restricted model (M_1) is estimated in which the tested factor loadings are constrained to be equal across groups and the fit between M_0 and M_1 compared. The tenability of the equality constraints is often tested statistically using the likelihood ratio test (LRT) within the framework of conventional null hypothesis significance test (NHST).

That is, under the assumptions of multivariate normality a true null hypothesis (i.e., the tested parameters are equal in the population), the chi-square difference (derived from likelihood ratio) between the two nested models (i.e. $T_{dif} = T_0 - T_1$) asymptotically follows a central chi-square distribution with degrees of freedom $df_{dif} = df_1 - df_0$ (Steiger, Shapiro, & Browne, 1985).

T_0 and T_1 represent the chi-square statistics for the baseline model and the more restricted model, respectively; df_0 and df_1 are the corresponding degrees of freedom. When the observed chi-square exceeds a critical value, determined by both df_{dif} and the alpha level (e.g. $\alpha = .05$), the null hypothesis of invariance is rejected (i.e., non-invariant parameter(s) are detected). On the other hand, if the chi-square difference test indicates non-significance, meaning the model with constraints fits the data as well as the baseline model, researchers would "accept" the constrained model and conclude that the tested parameters are invariant across groups.

Despite common use by applied researchers, the current practice of utilizing LRT to test measurement invariance is may be adversely affected by several phenomena. In this paper, we focus on two major issues that can affect the LRT. First, under the hypothesis testing framework, if the hypothesis test is found to be significant, the null hypothesis of invariant parameters is likely to be false. This is because the probability of making incorrect decision (i.e. Type I error) is controlled by the set alpha level (e.g. $\alpha = .05$). However, when the hypothesis test is non-significant, the result may simply be due to lack of power to reject the null hypothesis (i.e. Type II error). In other words, failing to reject does not provide any information regarding accepting the null hypothesis (Cohen, 1994). In application of LRT to testing factorial invariance, conventionally, the null hypothesis states that the tested invariance condition (i.e., the equality constraint imposed) holds in the population. Therefore, if results indicate a non-significant chi-square difference, one cannot confidently claim no cross-group differences on the tested parameters, unless the power rate is sufficiently high. In reality, however, the LRT could possess a relatively low level of power to identify non-invariant parameters, even when the magnitude of non-invariance is rather noticeable and the sample size is large. Such results have been demonstrated in the measurement invariance literature. For example, French and Finch (2006)

found the power of detecting a non-invariant factor loading with a cross-group difference of 0.25 was 51.3% (α=.05), even when using a moderate sample size of $N = 500$. The results imply that researchers would mistakenly conclude non-invariant factor loadings to be invariant in about half of the time if LRT difference testing is used. Erroneously considering non-invariant parameters to be equivalent could lead to some undesirable consequences, such as producing biased estimates when conducting cross-group comparisons on latent means and latent variances (French & Finch, 2016; Shi, Song, & Lewis, 2017).

Second, the LRT compares models with and without the imposed equality constraints, and asks if there is *no* difference between the models in terms of fit. In other words, the difference test is examining if the parameters of interest are *exactly* equal across groups. Consequently, as sample size increases, any cross-group differences in the tested parameters will yield statistically significant results. Thus, with sufficiently large samples, even if the non-invariance is minor and practically negligible, the hypothesis of factorial invariance is very likely to be rejected. Previous research has demonstrated this finding as well. For example, in the context of Item Response Theory (IRT) models, Meade (2010) showed that even trivially small levels of non-invariance[2] produced statistically significant LRT results when sample size reached 1,000.

Given that applying the conventional LRT to test factorial invariance is a pervasive practice, the aforementioned two issues likely arise frequently, leading to questionable conclusions. On the one hand, if a non-significant LRT is obtained (i.e., accepting the null hypothesis), researchers are at risk of mistakenly claiming truly non-invariant parameter(s) as invariant due to low power to detect the invariance. On the other hand, as sample size increases, the LRT would eventually yield statistically significance if any negligible level of non-invariance

exists. As a result, researchers may assert that a test does not measure the construct equivalently and abandon the measure for cross-group investigations, when, in fact, the level of non-invariance is practically negligible.

Methodological study has shed some light on overcoming the undesirable issues of LRT, and a few alternative tests to factorial invariance have been proposed. As a response to the aforementioned problems, Cheung and Rensvold (2002) recommended comparing practical goodness of fit indices (e.g., the comparative fit index, CFI) of nested models to test factorial invariance. Later researchers (Chen, 2007; Meade, Johnson, & Braddy, 2008) developed this approach further by evaluating and proposing the cutoffs to detecting non-invariance for a few commonly used fit indices, such as the comparative fit index (CFI) and the root mean squared errors of approximation (RMSEA). The approach of using change in fit indices (e.g., ΔCFI) to detect non-invariance has gained its popularity among empirical researchers. As of this writing, the three articles noted above have received nearly 10,000 citations[3]. Despite the popularity of using change in fit statistics to evaluate factorial invariance, there are also shortcomings in applying this approach. One main concern is that the approach based on fit indices is largely heuristic and is not grounded in statistical theory. That is, the procedures are conducted solely by evaluating the differences in the estimated fit indices (i.e., sample statistic), whereas the sampling variability of the statistic is blatantly ignored. Moreover, the cutoffs for determining non-invariance were generated from simulation results. As simulation conditions can greatly differ from empirical research situations, it is not easy to come up with reference points that can be applied in general settings. For example, Cheung et al. (2002), Chen (2007), and Meade et al. (2008) all recommend use of ΔCFI as the main criterion for testing factorial invariance; however, the three groups of researchers suggested different cutoffs for this difference[4].

Recently, Yuan and Chan (2016) introduced an equivalence testing approach for evaluating factorial invariance. Specifically, for assessing invariance, the null hypothesis of the equivalence test is set as H_0: $(F_0-F_1) > \varepsilon_0$. F_0 and F_1 are the values of the model fit function for the baseline model and the more restricted model, respectively; ε_0 is the maximum tolerable model misspecification (caused by non-invariance). If the null hypothesis is rejected, the researchers would conclude that the difference in model fit between the baseline model and the more restricted model is no larger than some small number ε_0 [H_1: $(F_0-F_1) \leq \varepsilon_0$]. In other words, by constraining the tested parameter(s) to be equal across groups, the parameters are considered equivalent if changes in model fit do not exceed an acceptable level of misfit (ε_0). To better define the acceptable level of misfit, Yuan et al. (2016) showed that the values of ε_0 can be linked to and interpreted on the metric of RMSEA, a widely used fit index in SEM. Therefore, the commonly used cutoffs for RMSEA (MacCallum, Browne, & Sugawara, 1996), as summarized below, can be applied:

Excellent fit: <.01
Close fit: .01-.05
Fair fit: .05-.08
Mediocre fit: .08-.10
Poor: >.10

For example, after adding equality constraints on factor loadings to the baseline model, if the equivalence test supports that the change of model fit is corresponding to the RMSEA of .02, the fit of the baseline model and the more restricted model is deemed “fairly close”. The researchers could gain statistical evidence and conclude that the weak invariance holds. However, it is noted that the conventional cutoffs for RMSEA are believed to be overly stringent for the purpose of assessing invariance. As such, Yuan and Chan (2016) proposed the formula and recommended to use the adjusted cutoff values for assessing factorial invariance[5].

The equivalence testing approach offers the potential to overcome the above-mentioned two problems of LRT in testing factorial invariance. Since the tests are conducted within the framework of equivalence testing (see Dunnett & Gent, 1977; Wellek, 2010; Yuan, Chan, Marcoulides, & Bentler, 2016; Marcoulides & Yuan, 2017), rejecting the null hypothesis could provide statistical evidence to support that factorial invariance holds. In addition, the equivalence testing approach explicitly informs researchers the size of model misspecifications (i.e., non-invariance). Thus, the level of non-invariance can be explained and statistically tested based on the metric of RMSEA, allowing researchers to evaluate the practical importance of the non-invariance. Recently, the equivalence testing approach for testing factorial invariance has been adopted in empirical applications (e.g., Testa et al., 2017; Contractor et al., 2018) and supported by a simulation study (Finch & French, 2018).

Although the equivalence testing makes an important contribution to the methodology of testing factorial invariance, there are a few limitations. First, as noted by Yuan and Chan (2016), due to the way the null hypothesis is set, Type I error and power under the equivalence testing framework have different implications from those under the conventional NHST. That is, if the ultimate goal is to endorse factorial invariance, the equivalence testing approach would provide statistical evidence to support the proposed invariant constraints with a low error rate (controlled by the alpha level). On the other hand, failing to reject the null hypothesis only implies that there is not enough evidence to endorse factorial invariance; researchers cannot confidently conclude non-invariance is detected, especially when the sample size is small (i.e., power is low).

Second, the equivalence testing approach allows researchers to quantify the level of misspecificaion (non-invariance) on the metric of RMSEA and thus to make more informative decisions. However, it is noted that the values of RMSEA is impossible to interpret because it is

in an unstandardized metric (Maydeu-Olivares, 2017; Shi, Maydeu-Olivares, DiStefano, 2018). As Edwards (2013, p. 213) puts it "We do not know what a 0.01 difference in RMSEA values means. We do not know that a model with an RMSEA of 0.12 is incapable of telling us something useful about the world. We do not know that a model with an RMSEA of 0.01 is telling us anything useful about the world." Besides the level of model misspecification, the RMSEA is dependent on other characteristics of the fitted model (i.e., "incidental parameters", Saris, Satorra, & van der Veld, 2009). For example, the same RMSEA (say 0.05) may hold a different meaning in terms of the model misspecification when models differ in terms of the magnitude of factor loadings and model size (Chen, Curran, Bollen, Kirby, & Paxton, 2008; Savalei, 2012; Maydeu-Olivares, Shi, & Rosseel, 2018). Therefore, for testing factorial invariance, the level of non-invariance may not be well communicated based on the metric of RMSEA.

Methodologists have proposed a few effect size indices for the purpose of better quantifying the magnitude of non-invariance. These proposed effect sizes have different meanings and can serve different purpose for evaluating the consequences of non-invariance (see Meade, 2010 for a review). For example, the effect size can be indicated by the standardized differences in the metric of the tested parameters (Steinberg & Thissen, 2006). An alternative effect size measure is to directly evaluate the consequences of non-invariance depending upon the specific use of the test scores (e.g. the impact of non-invariance on selecting individuals; Millsap & Kwok, 2004; Lai, Kwok, Yoon, & Hsiao, 2017). The usage of the effect size indices allows researchers to communicate the magnitudes of non-invariance in an interpretable way. However, the current practice of using effect size of non-invariance is mostly prestatistical. That

is, the effect sizes are reported and interpreted solely depend on the sample estimates; very few, if any statistical tests is available to make inference of the effect size parameters.

In the current study, we introduce an interval estimation approach to invariance testing based on Bayesian structural equation modeling (BSEM) to address the two noted shortcomings from LRT. The proposed approach quantifies the size of non-invariance in an interpretable manner, with an uncertainty interval evaluated using the Bayesian parameter estimation. Therefore, the Bayesian interval approach allows researchers to support the null hypothesis of practical invariance, and examine the practical importance of the non-invariant parameter. The article is organized into three sections. First, we present a detailed discussion of the Bayesian interval estimation approach. Next, the performance of the proposed method is evaluated using simulation. Finally, we provide an empirical example to demonstrate the use of the proposed method using data from a depression study.

Bayesian Assessment of Null Values Via Parameter Estimation

In recent years, Bayesian approaches have been increasingly applied in latent variable models (e.g. Lee, 2007; Lee & Song, 2012; Muthén & Asparouhov, 2012). Under a Bayesian framework, let **M** be an arbitrary SEM model with the unknown parameters in a vector $\boldsymbol{\theta}$, and let **Y** represent the observed data. A standard Bayesian approach requires the evaluation of the posterior distribution of $\boldsymbol{\theta}$ given **Y** (i.e., $\Pr(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{M})$). This can be obtained by $\Pr(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{M}) \propto \Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{M}) \Pr(\boldsymbol{\theta} \mid \mathbf{M})$ based on Bayes' Theorem, where $\Pr(\mathbf{Y} \mid \boldsymbol{\theta}, \mathbf{M})$ is the likelihood of observing data **Y** conditional on the parameters $\boldsymbol{\theta}$, and $\Pr(\boldsymbol{\theta} \mid \mathbf{M})$ is the prior probability of the parameters $\boldsymbol{\theta}$.

To analytically obtain a solution for the posterior distribution, $\Pr(\boldsymbol{\theta} \mid \mathbf{Y}, \mathbf{M})$, numerical integration would be used to obtain the posterior mean and posterior variance for each model

parameter. However, when the model involves latent variables and many parameters, the high-dimensional integration often has no closed form and consequently, the posterior mean and variance cannot be obtained analytically. Markov chain Monte Carlo (MCMC) methods can handle such otherwise intractable calculations. With MCMC, the basic idea is to repeatedly draw random numbers from a (full or conditional) posterior distribution and empirically summarize those draws (Martin, 2008, Gill, 2014). In Bayesian estimation of the measurement model in SEM, a data augmentation technique is used (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 can be obtained. Differing from a "traditional" statistical perspective (i.e., frequentist) which typically treats parameters as constants, the Bayesian framework treats model parameters as random variables. The parameter estimates are then obtained as the empirical means, modes, or medians of a posterior distributions (Song & Lee, 2012).

Statistical inferences may be made directly via Bayesian parameter estimation, where the credibility of the estimated parameter is obtained from its posterior distribution (Kruschke & Liddell, 2017; Kruschke, 2011). For each parameter of interest, the uncertainty of estimation can be captured by the highest density interval (HDI) given the representative observations from the posterior distribution. The highest density interval indicates which points (i.e., possible parameter values) are most credible. That is, parameter values within the HDI have higher credibility than the values lying outside the interval. A more formal definition of HDI is given as (Kaplan, 2014, Pg. 96):

Let p(θ|y)be the posterior probability density function. A region R of the parameter space θ is called the 100(1-α) % highest density interval if

1. $p(\theta \in R \mid y) = 1 - \alpha$

2. For $\theta_1 \in R$ *and* $\theta_2 \notin R$, $p(\theta_1 \mid y) \geq p(\theta_2 \mid y)$

In practice, the HDI can be interpreted in a probability manner. For example, it can be stated that there is a 95% chance that a parameter falls in the 95% HDI, which is generally believed to contain the most credible values of the parameter (given information from the data and the specified prior). The HDI could be used as a hypothesis testing decision tool. That is, if a tested null value (e.g., zero) is within the 95% HDI, the null value *is not* rejected; but, if the 95% HDI does not include the null value, the null value *is* rejected (Kruschke, Aguinis, & Joo, 2012).

In addition, by providing a range of parameter values that cover most of the posterior distribution, inferences regarding the practical effect size can be made. Specifically, a *region of practical equivalence* (ROPE) is predetermined by the researcher, which indicates "a small range of parameter values that are considered to be practically equivalent to the null value for purposes of the particular application" (Kruschke, 2014, pg. 336). For example, considering a correlation coefficient, a possible choice of ROPE is ±.10, because values within the region -.10 to .10 indicate a very weak correlation that may be considered practically equal to zero for some empirical applications. Statistical decisions can be made by comparing the researcher-determined ROPE with the 95% HDI. The null value for the tested parameter is retained if the ROPE determined by the researcher completely contains the 95% HDI, because all of the most credible parameter values are practically equivalent to the null. By using the similar logic, when the 95% HDI completely excludes values from the ROPE, the null value is rejected. Nevertheless, if the 95% HDI and ROPE partially overlap, neither of the above conditions are satisfied, and the research is proven inconclusive (Kruschke & Liddell, 2017; Kruschke, 2014).

As shown above, the Bayesian parameter estimation framework can be used as a tool for conducting statistical tests. By using ROPE, a Bayesian interval estimation approach could provide richer information on the practical effect size with a given uncertainty interval. This

approach has been applied in a few modeling settings, such as comparing group mean differences (Kruschke, 2013) and testing parameters in regression models (Kruschke, 2014). In this study, we utilize the Bayesian interval estimation approach in the context of a CFA model and thereby propose a new method for evaluating factorial invariance.

An Interval Estimation Approach to Testing Factorial Invariance using BSEM

When considering invariance testing in a Bayesian interval estimation approach, we first define a new parameter, d_{ij}, to represent the cross-group difference in a specific factor parameter, *i,* for an item, *j*. For example, $d_{\lambda 1}$ and $d_{\tau 1}$ denote the cross difference in the factor loadings (λ) and intercepts (τ) for item 1, such that

$$d_{\lambda 1} = \lambda_{1(1)} - \lambda_{1(2)} \quad (2)$$

$$d_{\tau 1} = \tau_{1(1)} - \tau_{1(2)} \quad (3),$$

with the numbers in parentheses representing group membership.

It is noted that the invariance tests typically compare factor model parameters on their raw metrics. Therefore, the meaning of d_{ij} can be ambiguous depending on the scales of the factors and/or the raw data. To avoid difficulty in interpretation, for both factor loadings and intercepts, the raw differences can be standardized using the pooled standard deviation across groups. The corresponding standardized parameters (D_{ij}) are used as measures of effect size for non-invariance.

For factor loadings, the standardized difference $D_{\lambda j}$ can be expressed as

$$D_{\lambda j} = d_{\lambda j} \times \frac{S_f}{S_{yj}}, \quad (4)$$

where S_f and S_{yj} denote the pooled standard deviations for the latent factors and observed variable (for item *j*), respectively, across groups. In the case of two groups, pooled standard deviation values may be calculated as:

$$S_f = \sqrt{\frac{(N_{(1)}-1)s^2_{f(1)} + (N_{(2)}-1)s^2_{f(2)}}{(N_{(1)}-1)+(N_{(2)}-1)}} \tag{5}$$

$$S_{yj} = \sqrt{\frac{(N_{(1)}-1)s^2_{yj(1)} + (N_{(2)}-1)s^2_{yj(2)}}{(N_{(1)}-1)+(N_{(2)}-1)}}\,, \tag{6}$$

where *N* represents the group sample size, s^2_f represents the factor variance and s^2_{yj} is the variance for the observed variable *j*. The standardized differences in the intercepts for item *j* ($D_{\tau j}$), using the same notation, can be expressed as:

$$D_{\tau j} = \frac{d_{\tau j}}{S_{yj}}, \tag{7}$$

where the standardization process only considers the scale of the observed variables.

Provided that the multiple-group CFA model are identified and scaled by using the correct metric[6], if the tested factor parameter is truly invariant across groups, the corresponding D_{ij} is zero in the population. For a non-invariant parameter, however, the population D_{ij} is non-zero and the values of D_{ij} inform the size of the non-invariance. Specifically, a positive D_{ij} implies that the non-invariant parameter is larger in the first group and a negative D_{ij} implies the opposite. Larger $|D_{ij}|$ values suggest larger differences across groups, and a more acute level of non-invariance. Furthermore, D_{ij} serves as a standardized effect size of non-invariance. That is, D_{ij} can be interpreted as the standardized difference of the tested parameters across groups, and the values of D_{ij} can be compared across items/tests and studies.

In fitting BSEM model, for each tested parameter, the corresponding D_{ij} can be introduced as a new random variable (following Equations 4 and 7 above). Assuming there is no prior evidence regarding invariance, non-informative priors are used for all estimated parameters. Using these priors, the posterior distributions, thus the 95% HDI for all D_{ij} can then be obtained simultaneously via Bayesian estimation. Therefore, the BSEM approach not only provides an estimate of the size of non-invariance, but also, the sampling errors are taken into consideration with an uncertainty interval.

By checking the posterior distributions and 95% HDIs for D_{ij}, researchers are allowed to quantify the size of non-invariance as a "continuum". However, from the applied perspective, a subjective decision making procedure is usually unavoidable, because eventually the researchers have to decide whether the test is useable (i.e., whether the tested parameter is practically invariant), or not. Using the BSEM approach, more informative decisions on factorial invariance can be obtained by incorporating the information from a region of practical equivalence (ROPE). Here, the ROPE indicates a range of values for D_{ij} that are considered to be practically ignorable, as predetermined by researchers. Specifically, by examining the relationship between the 95% HDI of D_{ij} and the predetermined ROPE interval, four possible decisions regarding invariance are obtained. Figure 1 summarizes the decision process using a flow chart. Specifically, four decisions may be made. First, if the 95% HDI for D_{ij} falls completely within the ROPE interval, the tested parameter is concluded to be practically invariant. Second, if the 95% HDI for D_{ij} falls completely outside the ROPE interval, significant non-invariance is identified and the detected non-invariance is practically important. Third, if the 95% HDI does not include zero, and partially overlaps with the ROPE, the tested parameter is significantly non-invariant, but no sufficient evidence can be provided regarding the practical importance of the non-invariance

(vis-à-vis the selected ROPE). Finally, if the 95% HDI includes zero and it partially overlaps with the ROPE, the result is inconclusive, and researchers should refrain from stating a decision.

It is noted that the selection of ROPE can be somewhat subjective but unavoidable. A direct analogy is that when assessing SEM models, researchers could objectively compute the RMSEA (and the confidence interval) based on the statistical theory; however, if a decision must be made regarding the model fit, an agreed upon cut-off value has to be selected subjectively. We argue that the choice of ROPE should be based on the notion of the substantively ignorable non-invariance. Specifically, in this paper, we employ the standardized parameter difference (D_{ij}) to measure the size of non-invariance. Thus, we define a standardized difference (D_{ij}) as substantively ignorable if applied researchers would retain the restricted model (where the tested parameters are fixed to be equal across groups) should they know what the true model is. For instance, most researchers would agree that a standardized factor loading of .10 is small; therefore, a model constraining two factor loadings with a standardized difference of .10 to be equal is considered acceptable (although it is misspecified). In contrast, most researchers would not fix the factor loadings to be equal should they know the standardized factor loading difference is .40. As a result, a ROPE for D_{ij} with limits of ±.10 should be appropriate, not for D_{ij} with limits of ±.40. In this study, we used two ROPEs for D_{ij} with limits of ±.10, and ±.20. The two selected ROPEs represent a relatively strict (±.10) and a more liberal criterion (±.20) for practical invariance.

It is also worth mentioning that in this study, the effect size of non-invariance was measured using the standardized parameter difference (D_{ij}) for two considerations. First, the standardized parameter difference (D_{ij}) could quantify the size of non-invariance for each model parameter in an interpretable manner. Second, as the definitional function is simple, the

posterior distributions for D_{ij} is easy to obtained using the user-friendly software (e.g., Mplus), even for applied researchers without much training in Bayesian statistics or programing. Nevertheless, the Bayesian interval estimation approach described above can also be applied with other effect size measures for non-invariance (Meade, 2010), depending on the purpose of the measure. We will revisit this point in the discussion.

Monte Carlo Simulation

We examined the performance of the proposed approach in evaluating factorial invariance through a simulation study.

Data Simulation

Multivariate normal data were generated based on a multiple-group CFA model. We restricted the number of groups to two, and for each group, five items loaded on a common latent factor. In both groups, the population factor mean and factor variance are set to be zero and one, respectively; the error variances were set such that the standard deviations of all observed variables equaled 1.0.

In group 1, all factor loadings were set to 0.80, and all intercepts were set to 0. In group 2, the first four items are invariant with factor loadings and intercepts equal to those of the first group (i.e. factor loadings equal to 0.8 and intercepts equal to 0). Possible non-invariant parameters are only manipulated in item 5; the population values for the non-invariant factor loading and intercept in group 2 are determined according to different simulation scenarios. Other characteristics that were manipulated are as follows:

Sample size. The two groups were generated with equal number of observations. Sample sizes include 100, 200, 500, 1,000, 2,000, and 10,000 per group. The levels were chosen to represent relatively small to very large samples in social sciences.

Source and Magnitude of non-invariance. Non-invariance was simulated either on the factor loadings or intercepts. Four levels of non-invariance were considered: invariant, trivial, small and large non-invariance. The invariant conditions imply no cross-group differences on the factor model parameters (i.e. factor loadings and intercepts). For conditions with negligible cross-group difference, factor loadings in the second group decreased by 0.05, or intercepts increased by 0.05. Small cross-group differences occurred when the factor loadings in the second group decreased by 0.2, or intercepts increased by 0.3. Under the large difference conditions, factor loadings in the second group decreased by 0.4, or intercepts increased by 0.6. The choices for the magnitudes are based on suggestions from previous literature (Kim, Yoon, & Lee, 2012; Kim & Yoon, 2011; Meade & Lautenschlager, 2004).

The manipulated variables were fully crossed, 6 sample sizes × 2 source of non-invariance × 4 magnitude of non-invariance, for a total of 48 conditions. For each simulation condition, 500 replications were generated and analyzed with Mplus 7.11 (Muthén & Muthén, 1998-2012).

Data Analysis

A Bayesian multiple-group CFA model was fit to each simulated dataset. For model identification, the first group is used as the reference group and its factor mean and factor variance were fixed to be zero and one, respectively. The first item was selected as the reference indicator, in which the factor loading and intercept for item one was constrained to be equal across groups. All other parameters are freely estimated. Invariance was only evaluated for the factor model parameters (i.e. factor loading or intercept) on item 5, where the corresponding D_{ij} was defined as a new variable. Non-informative priors were used for all estimated parameters. By default, Mplus sets the priors of factor loadings, intercepts and the factor means to Normal (0,

infinity); the priors for the residual variances and factor variances are set to be to Inverse-Gamma (-1, 0). Since these settings are widely spread and therefore contain little information about the distributions of the parameters, they are regarded as non-informative.

Two MCMC chains were utilized; each chain had 100,000 iterations where the first half of iterations were discarded as burn-in. As a result, the final posterior distribution for each estimated parameter was constructed from a total of 100,000 draws. There are multiple ways to assess the convergence status of the MCMC chain. Perhaps the most popular one is Potential Scale Reduction (PSR; Gelman & Rubin, 1992) which is obtained via diving the between-chain variation by the total variation. When the PSR reaches to one, it indicates that the multiple chains have converged to the same distribution and therefore the mixing process is sufficient. Mplus provides the maximum PSR value of all parameters across iterations for the diagnosis of convergence. Through a pilot investigation of 20 replications for each simulated conditions, it is found that the maximum PSRs are be guaranteed to fall under 1.001 after approximately 10,000 iterations. Given the iteration number was set to 100,000 throughout the replications, it can be trusted that the posteriors were yielded at a converged condition. In addition to the PSR, other methods were proposed to assess Markov Chain convergence, for example, Geweke (1992) compares means calculated from distinct segments of Markov chain, where Raftery and Lewis (1992) estimates the minimum chain length needed to estimate a percentile to some precision. Details about the convergence diagnosis can be found in Alkan (2017).

To evaluate factorial invariance, the posterior distributions of D_{ij} for the tested parameter was examined. Flowing the default in Mplus, the median of the posterior distribution is used as the point estimate, and the 95% HDI was also obtained. For each simulation condition, we compute the average point estimates for D_{ij}, and compared them with the population values. The

population coverage rates of the 95% HDI were also obtained, which is defined as the probability of the 95% HDI including the population value of D_{ij}.

For making practical decisions, the 95% HDI was further compared with the ROPE interval. Two different ROPEs were considered with limits of ±.10, and ±.20, respectively. For each replication, we followed the rule discussed in the previous section, that is, a decision for invariance was made by considering the four possible outcomes of the Bayesian interval estimation approach.

We also performed the factorial invariance tests using the LRT, the goodness of fit difference test (ΔCFI), and the equivalence test. Specifically, we compared the baseline model (where equality constraints are only added on the reference indicator for identification purpose) and the constrained model (where equality constraint is added to each tested parameter). Under LRT, the tested parameter was either invariant (e.g. non-significant LRT) or non-invariant (significant LRT). Using the goodness of fit difference test, |ΔCFI ≥.01| indicated non-invariance; whereas |ΔCFI <.01| implies invariance[7]. For the equivalence testing approach, we adopted the adjusted cutoff values proposed by Yuan and Chan (2016), and the size of non-invariance is categorized into five levels of fit based on the metric of RMSEA. The five levels of fit included excellent, close, fair, mediocre, and poor. According to Finch and French (2018), either excellent or close fit indicated practical invariance, and the category of poor fit generally implied practical non-invariance. For the three alterative tests described above, maximum likelihood (ML) estimation was used. The LRT and ΔCFI tests were conducted using Mplus; the equivalence testing was conducted with the R (R Core Team, 2015) using the code provided in the appendix of Yuan and Chan (2016). To evaluate and compare the performance among

different approaches, for each simulation condition, the empirical probability of making all possible decisions was computed across the 500 replications.

Results

Table 1 summarizes the average point estimates and the coverage rates for the 95% HDI around the population value of D_{ij}. The results showed that across all simulated conditions, the average point estimates were fairly close to the population values. In addition, the 95% HDI performed well in terms of covering the true effect size (D_{ij}), yielding coverage rates which ranged from 0.93 to 0.97. Therefore, by examining the posterior distribution, the BSEM approach could provide the unbiased estimation of the effect size of non-invariance (D_{ij}) with an accurate uncertainty interval.

We further investigated the performance of using the BSEM approach to make practical conclusions regarding factorial invariance, and the performance was compared with the three existing methods (i.e., the LRT, the Δ CFI test and the equivalence testing approach). In Table 2 and Table 3, we summarize the probabilities of making different conclusions for evaluating invariance on factor loadings and intercepts, respectively.

The results suggested that when using the LRT approach, researchers would make misguided conclusions under certain conditions. That is, in small samples, by groundlessly claiming the parameters with non-significant chi-square differences are invariant, the LRT approach was likely to lead to relatively high Type II errors. For example, when $N = 200$, there was a 48% chance that researchers would conclude a non-invariant factor loading when there was a true cross-group difference of 0.20. Even when the magnitude of non-invariance was very large ($D_{\lambda} = .40$), researchers still have 17% chance to conclude that the tested factor loading was invariant if the sample size $N = 100$. On the other hand, with a sufficiently large sample size, the

power of LRT approached one; any non-invariant parameter was eventually rejected by LRT, even if the non-invariance was trivial. For example, as N increased to 10,000, the power rate for rejecting the null with an intercept with trivial non-invariance (e.g. D_τ= -.05) was 0.98.

As discussed above, the LRT was highly sensitive to sample size. The Δ CFI tests overcame this issue of the LRT. It is noted from Tables 2-3 that when the tested parameter was truly invariant, or the magnitude of non-invariance was trivial, by using the Δ CFI tests, researchers would have almost zero probability to conclude the parameter was non-invariant, even the sample size was very large (e.g., N = 10,000). However, in the presence of nonignorable non-invariance, the Δ CFI tests can be underpowered. That is, the Δ CFI tests tended to fail to detect the non-invariant parameters, especially when sample size was small. For example, using the Δ CFI tests, when N = 200, the probability of mistakenly concluding that factor loadings with small (D_λ= .20) and large (D_λ= .40) non-invariance as invariant were 92% and 34%.

Under the equivalence testing approach, researchers are allowed to quantify the size of non-invariance into five categories based on the metric of RMSEA. As we can see from the tables, the equivalence testing approach generally performed well in making accurate conclusions, especially when the sample size was larger than 500. Specifically, when $N \geq 500$, using the equivalence testing approach, the probability of concluding truly invariant parameters or parameters with trivial non-invariance as practically invariant (i.e., under the columns of excellent fit and close fit) was larger than 70% across all simulated conditions. Moreover, the probability to detect noticeable non-invariance (i.e., under the column of poor fit) was generally greater than 80% ($N \geq 500$). When sample size was small (i.e., $N < 500$), by applying the equivalence testing approach, researchers might erroneously conclude truly invariant parameters or parameters with trivial non-invariance as practically non-invariant. For example, when N =

100, the probability to conclude that the truly invariant factor loading (D_λ= .00) and the trivially non-invariant factor loading (D_λ= .05) to be practically non-invariant were 19% and 25%.

Four outcomes can be derived by applying the BSEM approach for testing invariance: 1) practically invariant, 2) importantly non-invariant, 3) non-invariant with an uncertain practical importance (vis-à-vis the selected ROPE), and 4) inconclusive. Results showed that the probability of making an erroneous conclusion was very low for the BSEM approach across all simulated conditions. Specifically, the probability of concluding a truly invariant parameter to be non-invariant (i.e., either importantly non-invariant or non-invariant with uncertain practical importance) was less than 6%. The BSEM approach also almost never suggested a trivial non-invariance to be practically important, even in very large samples (e.g. N=10,000). Further, when the non-invariance was noticeably large, the BSEM approach yielded almost zero chance to "accept" such a parameter as practically invariant.

In addition, based on the relationship between the 95% HDI and the region of practical equivalence (ROPE), the BSEM approach could offer additional insight into evaluating factorial invariance, especially when sample size was larger than 500. That is, if the most credible values of non-invariance (i.e. 95% HDI) were completely within the ROPE, the BSEM could provide direct evidence to support that the tested parameter is practically invariant. For example, as shown in Table 1, when $N \geq 500$, using a ROPE with limits ±.20, the probability of "accept" the truly invariant factor loading or trivially non-invariant factor loading (as practically invariant) was larger than 70% across all simulated conditions. When some noticeable level of non-invariance was present, the BSEM approach incorporated effect size information into the decision, allowing researchers to determine whether the non-invariance is practically important. For example, if using a cutoff of ±.20 to denote practically important non-invariance, when $N \geq$

500, the BSEM approach had no less than 87% chance to identify the factor loading with D_λ= .40 as importantly non-invariant. However, with the same ROPE (±.20) and sample sizes ($N \geq 500$), if the standardized factor loading difference D_λ= .20, which was right on the cutoff of ROPE; for most of the time (with probabilities larger than 86%), the BSEM approach concluded that the non-invariance was statistically significant, but researchers cannot determine its practical importance.

It was not surprising to observe that for a given level of non-invariance, the conclusion was dependent upon the choice of ROPE, as well as the sample size. When the ROPE had narrower limits, more evidence (i.e., a larger sample) was required to confidently support a stricter criterion for practical invariance. For example, given truly invariant factor loadings, by applying ROPE with limits ±.10, the BSEM approach could not provide support for practical invariance when the sample size was small (i.e., $N \leq 500$), and instead of making unfounded conclusions, inconclusive decisions were generated almost every time. As N increased to 2,000, the probability of concluding the tested factor loading illustrated practical invariance (i.e. $|D_{\lambda j}|<.10$) increased to 86%. Smaller samples, however, can offer sufficient support for a more liberal definition of practical invariance. For example, when N = 500 and a wider ROPE with limits ±.20 was used, a truly invariant factor loading was accepted as practically invariant with a probability of 82%.

On the other hand, if the goal was to reject the non-invariant parameters, for a fixed level of non-invariance, a narrower ROPE yielded a higher probability of concluding the non-invariance was practically important, especially when N was small. For example, when the cross-group difference in the intercept was .30, N=100, and using ±.1 as the limits for the ROPE, the probability of concluding the non-invariant parameter practically important was 0.31. As the

ROPE became wider (e.g. with limits ±.20), the probability of concluding the same parameter importantly non-invariant decreased to 10%. As we demonstrated earlier, in an extreme case, where the magnitude of non-invariance is exactly at the limit of the ROPE (e.g. when the cross-group difference on the factor loadings was .20, and a ROPE with limits ±.20 was used), a researcher could confidently claim the non-invariance is statistically important, but she/he cannot determine whether such non-invariance is important (i.e., beyond the ROPE). However, if the level of non-invariance was noticeably further from the limits of ROPE (i.e. cases with large differences), and the sample size was relatively large (e.g. $N \geq 500$), the probability of claiming the tested parameters were importantly non-invariant was generally greater than 90%, regardless of choice for ROPE. It is also noted that when sample size was greater than 200, the probability of concluding non-invariance under the BSEM approach (i.e. either important non-invariance or non-invariance with uncertain practical importance) was approximately equal to the power rates from the LRT approach.

Discussion

In this study, we introduced a Bayesian interval estimation approach for evaluating factorial invariance. The standardized cross-group difference for each of the tested parameters with a highest density interval (HDI) can be obtained via Bayesian parameter estimation. Therefore, researchers may evaluate the invariance for each tested parameter using a single BSEM model. Decisions about invariance are made by examining the relationship between the 95% HDI and a region of practical equivalence (ROPE) which is selected by the researcher.

Using a simulation study, we compared the performance of the BSEM approach with three exciting methods. As compared to the LRT and Δ CFI test, which makes invariant/non-invariant decision, the BSEM approach provided richer information with the effect size of non-

invariance. Specifically, the Bayesian approach allows researchers to 1) support the null hypothesis of practical invariance, and 2) examine the practical importance of the non-invariant parameter. Therefore, under certain conditions, the BSEM approach could lead to more informative conclusions, which cannot be addressed when using the "traditional" approach to invariance testing.

In addition, results indicated that the BSEM approach showed comparable performance to the recent developed equivalence testing approach when the sample size was large ($N \geq 500$). That is, both approaches could yield similar probability to "accept" practical invariance and detect parameters that are noticeably non-invariant. When sample size was small ($N < 500$), the equivalence testing approach might erroneously conclude the truly invariant parameters and parameters with trivial non-invariance as practical non-invariant (i.e., poor fit). As we discussed earlier, these observations are due to the way the null hypothesis is set under the equivalence testing. Since the null hypotheses state that the tested parameters are practically non-invariant, failing to reject the null hypothesis only implies that there is not enough evidence to support invariance; researchers cannot confidently conclude non-invariance is detected, particularly when the sample size is small (i.e., power is low). We noted that the above observations do not imply that the equivalence testing was wrong; as the ultimate goal of the equivalence testing approach is to endorse factorial invariance (Yuan & Chan, 2016). In small samples ($N < 500$), the Bayesian interval estimation approach, on the other hand, tended to generate more conservative decisions. That is, if the evidence is not sufficient, the Bayesian approach proves inconclusive and suggests that the researcher should withhold a decision.

Other differences are noted between the BSEM approach and the equivalence testing approach. First, the equivalence testing approach quantifies the size of non-invariance in terms of

the metric of RMSEA; whereas the BSEM approach uses the standardized parameter differences as the effect size measure, which is more interpretable. Second, the proposed BSEM method only works when the number of groups (j) is two. The equivalence testing approach can be applied for invariance tests with many groups ($j \geq 3$). Finally, to properly apply tests for factorial invariance, the baseline model must be correctly, or closely specified (Yuan & Bentler, 2004; Maydeu-Olivares & Cai, 2006), the close fit of the baseline model can be tested using equivalence testing (see Yuan, Chan, Marcoulides, & Bentler, 2016; Marcoulides & Yuan, 2017; Yuan & Chan, 2016). In practice, we recommend researchers to select the approach based on the specific test situation they have (e.g., using the equivalence testing approach when the number of groups is larger than three). Generally speaking, the equivalence testing approach and BSEM approach showed similar performance when $N \geq 500$. In small samples ($N < 500$), the BSEM approach can be more conservative, and has less chance to draw erroneously conclusion by suggesting inconclusive evidence. In addition, researchers could apply one method to test factorial invariance and check the results by practicing another approach. More confident conclusion can be made if both approaches agree.

In applications of the BSEM approach, one crucial step is to select a ROPE. According to Kruschke (2014), "the [choice of] ROPE limits … cannot be uniquely 'correct', …and the limits of the ROPE depend on the practical purpose of the ROPE" (pg. 338). For evaluating factorial invariance, we recommend researchers to choose a ROPE based on the notion of the substantively ignorable non-invariance. In the current study, the effect size of non-invariance is measured by the standardized parameter difference (D_{ij}). We used two ROPEs for D_{ij} with limits of ±.10, and ±.20. The two selected ROPEs represent a relatively strict (±.10) and a more liberal criterion (±.20) for practical invariance. These choices of ROPEs and the obtained results could

be used as reference points empirical researchers, but it is noted that these ROPEs for practical invariance do serve as a fixed criteria. We acknowledge that this definition of substantively ignorable non-invariance is subjective, but we believe that this is how it should be, as the choice of ROPE on D_{ij} reflects how researchers define practical invariance, and the tolerance level of non-invariance depends on the purpose of the analysis. In practice, researchers may differ in what they consider a practically invariant parameter, depending on factors such as the construct under study, the purpose for using the test results, and the substantive theory related to the nomological network.

It is also worth noting that decisions made by the Bayesian approach may be affected by both the limits of the ROPE and the sample size. A narrower limit of ROPE implies a stricter criterion for practical invariance, thereby more evidence (i.e. a larger sample) is needed. Based on the results from the current study, roughly speaking, to confidently support an invariant parameter with ROPE limits $D_{ij} = \pm.10$, a sample of 2,000 observations is required.

When the size of non-invariant parameter is very close or equal to the one of the limits for the selected ROPE, it is difficult for researchers to determine whether such non-invariance was important (i.e. beyond ROPE) or not, even with large samples. Under such scenarios, in addition to selecting a "final" decision from the four available options (e.g., the tested parameter is non-invariant with uncertain practical importance), we recommend researchers always report and interpret the 95% HDI, to communicate the specific credible interval for the detected non-invariance.

A Pedagogical Example

we provide an empirical example to demonstrate evaluating factorial invariance using the Bayesian interval estimation approach using items from the Center for Epidemiologic Studies

Depression Scale (CES-D, Radloff, 1977). Data were obtained from the China Family Panel Studies, a nationally representative longitudinal survey conducted by the Institute of Social Science Survey of Peking University (funded by 985 Program of Peking University, Xie & Hu, 2014). Information from the 2012 wave of data was utilized and only participants who responded to all 15 items were included in the analysis (N=31,235). The average age of participants was 45.24 years (SD=16.64 years). Males made up approximately 48.83% of the sample and females comprised 51.17% of the sample.

Subjects are asked to utilize the CES-D to indicate how often they have felt depression symptoms during the past week. Responses are made on a four-point Likert-type scale ranging from zero (i.e., "Rarely or none of the time/Less than one day") to three (i.e., "All of the time/5-7 days"). We recognize that since the number response categories was small (i.e., less than five), to better account the ordinal nature of the data; ordinal factor analysis models (or polytomous IRT models) should be used (DiStefano & Morgan, 2014; Rhemtulla, Brosseau-Liard & Savalei, 2012). For demonstration purpose, here, we treated the outcome variables as continuous and fitted the ordinary CFA models in the example. The original version of the scale contains 20 items. For demonstration, a unidimensional set of 15 items was included in the analysis (see Edwards, Cheavens, Heiy, and Cukrowicz, 2010 for more information about the 1-factor structure). The 15 CES-D items used are provided in Table 4.

The factorial invariance test was conducted on the shortened version of CES-D across gender (i.e., male and female). In fitting the multiple group CFA model, females were used as the reference group, and their factor mean and factor variance were fixed to be zero and one, respectively. Item 3 (i.e., "I felt that I could not shake off the blues even with help from my family or friends") was selected as the reference indicator, following the procedures proposed by

Shi, Song, Liao, Terry, and Snyder (2017). For model identification, the factor loadings and intercepts for item 3 were constrained to be equal across genders. All other parameters were freely estimated.

The multiple-group BSEM model was fit by using Mplus commands "TYPE=MIXTURE" and "KNOWNCLASS" (see Muthén & Asparouhov, 2012 for details). Following Equations 1 and 2, the difference measure (D_{ij}) was defined as the difference between the same parameters across groups on a standardized metric using the keyword "NEW" under the "MODEL CONSTRAINT" option. The Mplus default non-informative priors were used for all estimated parameters. The priors of factor loadings, intercepts and the factor means are set to be Normal (0, infinity), and the priors for the residual variances and factor variances are set to be to Inverse-Gamma (-1, 0). For each of the two MCMC chains, 100,000 iterations were generated using the "FBITERATIONS" option. Under the "OUTPUT" command, "CINTERVAL (HPD)" was used to construct the highest density intervals for all model parameters; the "TECH8" option gave the progression of PSR values for assessing convergence diagnostics for the MCMC sampling. The "BPARAMETERS" option under "SAVEDATA" stored the Bayesian posterior parameter values from each iteration.The complete Mplus syntax for the multiple-group BSEM model is provided as supplementary material.

The convergence status for the MCMC sampling was accessed by the PSR. The "TECH8" output showed the PSR fall consistently under 1.005 after 50,000 iterations, implying that the posteriors were yielded at a converged condition. In addition, we asked the trace plots (plots of sampled MCMC values against iterations) for each estimated parameter using "TYPE=PLOT2" under the "PLOT" command. The trace plots are useful to evaluate the stationarity of the marginal posterior distribution. For item one, the trace plots of the standardized factor loading difference

($D_{\lambda 1}$) and the standardized intercept difference ($D_{\tau 1}$) were shown in Figure 2 and Figure 3, respectively. As we can see from the trace plots, the sequences for the difference parameters (D_{ij}) converged rapidly and the parallel chains mixed well together.

Factorial invariance was examined for all tested parameters simultaneously by checking the posterior distributions of the corresponding difference measures (D_{ij}). For all tested factor loadings and intercepts, their posterior distributions (with corresponding 95% HDIs and the ROPEs) were plotted in Figures 4 and Figure 5 using the BEST package in R (Kruschke & Meredith, 2015; R Development Core Team, 2015). The mean of the posterior distributions, as well as the 95% HDIs are reported in Table 5. In this example, we consider a relatively strict criterion for practical non-invariance, thus, ±.10 was used as the limits of the ROPE.

As seen from Table 5 and Figure 3, all tested factor loadings were found to be practically invariant across gender, except for items 14 and 17. The cross-group difference of factor loading for item 17 (i.e., "I had crying spells") was practically important ($\hat{D}_{\lambda 17} = .206$; 95% HDI= [.180, .234]). The large positive D implies that the association between depression and item 17 was noticeably higher for females than males. The factor loading in item 14 (i.e., "I felt lonely") illustrated significant non-invariance; but, using the set ROPE limits (i.e. ±.10), we cannot confidently claim the practical importance of the non-invariance ($\hat{D}_{\lambda 14} = -.079$; 95% HDI= [-.106, -.052]).

As shown in Table 5 and Figure 4, five items were found to have practically important non-invariance on the intercepts. Specifically, females exhibited noticeably larger intercepts than males for item 10 (i.e., "I felt fearful"; $\hat{D}_{\tau 10} = .147$; HDI=[.122, .171]), item 11(i.e., "My sleep was restless"; $\hat{D}_{\tau 11} = .127$; HDI=[.102, .151]), and item 17 (i.e.,. "I had crying spells";

$\hat{D}_{\tau 17} = .298$; HDI=[.275, .322]). On the contrary, the intercepts for item 9 (i.e., "I thought my life had been a failure"; $\hat{D}_{\tau 9} = .-125$; HDI= [-.151, -.100]) and item 13 (i.e., "I talked less than usual"; $\hat{D}_{\tau 13} = -.159$; HDI= [-.184, -.134]) were noticeably smaller for females than for males.

Item intercepts for item 14 (i.e., "I felt lonely"; $\hat{D}_{\tau 14} = -.092$; HDI=[-.117, -.065]), item 18 (i.e. "I felt sad"; $\hat{D}_{\tau 18} = .091$; HDI=[.065, .116]), and item 19 (i.e. "I felt that people disliked me."; $\hat{D}_{\tau 19} = -.098$; HDI=[-.124, -.072]) were significantly non-invariant across genders. However, no sufficient evidence was provided to claim whether the non-invariance is beyond the limit of ROPE (i.e.±.10). All other intercepts (i.e. Items 1, 2, 5, 6, 7 and 20) were found to be practically invariant.

Future directions

A few possible extensions of the proposed BSEM approach and future research directions are discussed. In the current study, the effect size of non-invariance was measured and interpreted using the standardized parameter difference (D_{ij}) cross-group (Steinberg & Thissen, 2006). For evaluating the practical importance of non-invariance, methodologists have developed and proposed a few other effect size indices (see Meade, 2010; Nye & Drasgow, 2011). For example, the effect size can be indicated by the cross-group differences on the expected observed (raw) scores, either at the item level or the test (or scale) level (Meade, 2010; Nye & Drasgow, 2011; Stark, Chernyshenko, & Drasgow, 2004; Nye, Bradburn, Olenick, Bialko, & Drasgow, 2018). An alternative effect size measure is to evaluate the consequences of non-invariance depending upon the specific use of the test scores (e.g. the impact of non-invariance on selecting individuals; Millsap & Kwok, 2004; Lai, Kwok, Yoon, & Hsiao, 2017). According to Meade (2010, pg. 730), different effect size measures could provide "slightly different, if generally

overlapping, information about the magnitude and nature of the [non-invariance].", and researchers can "get a fuller understanding of the [non-invariance] present by examining several indices rather than any one index alone". One avenue of future investigation is to apply other measures of effect size to the Bayesian interval estimation approach, and gain additional insight into understanding the practical importance of non-invariance.

In addition, an important feature of the Bayesian statistics is the use of priors (MacCallum, Edwards & Cai, 2012). Although many works have proved that Bayesian approach can improve the estimation accuracy of SEM (Hox, van de Schoot, & Matthijsse, 2012; Kim, Suh, Kim, Albanese, & Langer, 2013; Depaoli, 2014), inappropriate prior specifications, especially the ones without sound supports, can lead to inaccurate results (Depaoli, 2013; Depaoli, Yang, & Felt, 2017; Depaoli, 2013; Shi & Tong, 2017). As a safe strategy, which has been widely adopted in Bayesian studies, when no pre-judgment or educated guessing is available, non-informative priors are specified such that the posteriors can reflect as best as possible the information about the parameters estimated through the data. When the sample size of a dataset is sufficiently large, that priors being informative/non-informative produces less influence on the posteriors. On the other hand, a low sample size of studies can be sensitive to priors, particularly the ones being informative. Therefore, to investigate the robustness of Bayesian estimations, sensitivity to the prior parameter specifications should be tested when it is possible. In the present study, we assumed no prior evidence, therefore, non-informative priors were used for all estimated parameters. Since such priors carry little or no information about the parameter, estimation of parameters was predominately determined by the data. In applications, useful prior knowledge in terms of the presence and the size of non-invariance may be available based on results from previous studies (Kaplan & Depaoli, 2012; Muthén & Asparouhov, 2012;

Zondervan-Zwijnenburg, Peeters, Depaoli, & Van de Schoot). Therefore, by using informative priors, the Bayesian framework could offer methods for combining or incorporating evidence of factorial invariance from multiple studies. This possible extension for using BSEM to study factorial invariance deserves more attention.

Under the Bayesian framework, statistical inference can be conducted in several different ways (Kruschke, 2011a). In this study, we focused on the interval estimation approach using the posterior distribution. It is worth mentioning that within the Bayesian framework, some alternative approaches are available, and can possibly offer direct evidence for supporting factorial invariance. For example, traditionally, Bayesian hypothesis testing is based on a model comparison approach. Here, the Bayes factor is computed. The Bayes factor conveys the evidence of one model (e.g., M_0) as compared to another (e.g., M_1) given the data (D). This ratio compares and quantifies the evidence in favor of or against the null model (Dienes, 2011; Kruschke, 2011). For example, suppose the Bayes factor represents the ratio of the evidences in favor of the null model as compared to the alternative: $BF = \frac{\Pr(D \mid M_0)}{\Pr(D \mid M_1)}$. If the Bayes factor is substantially large (e.g. BF $\geq$ 3.0), the null model is considered better than the alternative, thus supporting the null hypothesis. If substantial evidence is found in favor of the alternative model (e.g. BF $\leq$ 1/3), the null hypothesis is rejected, and the alternative hypothesis is supported. 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. We expect future investigations to compare the performance between the proposed Bayesian interval estimation approach and the method based on Bayes factors in the context of testing factorial invariance.

Finally, the core idea of the interval estimation approach is to obtain the effect size of non-invariance (D_{ij}) with an accurate uncertainty interval. The interval estimation approach can

be pursued within the Frequentist approach as well (Falk, 2018). Specifically, the confidence interval (CI) for D_{ij} can be possibly constructed using the maximum likelihood with robust corrections to standard errors (Satorra & Bentler, 1988, 1994), or the nonparametric bootstrap (Efron & Tibshirani, 1993; Bollen & Stine, 1990). The *100(1-α) %* CI is defined as the interval that contains a population parameter (θ) *100(1-α) %* of the time, if researchers would use the same sampling method to select different samples and computed the interval estimate for each sample. It is noted that the interpretations of the Frequentist CI and Bayesian HDI (or Bayesian Credible Interval) are fundamentally different, although under certain conditions, using different intervals can provide similar results (Gelman, Stern, & Rubin, 2004). Future studies should further explore the possibility to use the Frequentist CI for testing factorial invariance and compare the results with the proposed Bayesian approach.

In summary, the Bayesian interval estimation approach could offer additional insight into understanding factorial invariance. The richer information gained through the Bayesian approach may lead to more insightful decisions about (non)invariance. We hope that this approach can assist applied researchers to make more accurate decisions when conducting invariance tests.

.

Footnotes

1. Different procedures have been developed for conducting LRT. In the current study, we focused on the procedure in which researchers first select at least one item as a reference indicator, and fit the baseline model with all other parameters freely estimated. Then, factorial invariance tests are conducted by fitting a series of models by imposing increasingly restrictive equality constraints (i.e., the free baseline approach). Alternatively, one can begin such tests by fitting a model with all of the parameters constrained to be equal, and then progressively relaxing certain equality constraints (i.e., the constrained baseline approach). Further information on the constrained baseline approach can be found in Stark, Chernyshenko, and Drasgow (2006) and Kim and Yoon (2011). In addition, non-invariance can also be detected by applying the iterative procedures (Cheung & Rensvold, 1998), in which each single item serves, in turn, as an RI (see also Cheung & Lau, 2012).

2. i.e., the difficulty parameter (*b*) decreased by .1 in Group 2.

3. Using Google scholar, up to 04/09/2018, the number of citations of Cheung et al. (2002), Chen (2007), and Meade et al. (2008) are 6595, 2125 and 584.

4. Cheung et al. (2002) recommended that $|\Delta \text{CFI}| \geq .01$ implied non-invariance; Meade et al. (2008) suggested that $|\Delta \text{CFI}| \geq .002$ implied non-invariance. Chen (2007) proposed cutoffs based on sample size; that is, the cutoffs for non-invariance were $|\Delta \text{CFI}| \geq .005$ for $N \leq 300$, and $|\Delta \text{CFI}| \geq .01$ for $N > 300$.

5. The revised cutoffs are larger than the conventional cutoffs for RMSEA, and are functions of the number of groups (*m*), sample size (*n*), and degrees of freedom (*df*). The computation details see Yuan and Chan (2016).

6. When fitting a CFA model, the metric of the latent variables must be set to identify the model. In testing for factorial invariance, a common method for identification is to use (at least) one item as a reference indicator (Cheung & Rensvold, 1999; Steiger, 2002; Johnson, Meade, & DuVernet, 2009). Specifically, an arbitrary group is selected as the reference group and its factor variance to set to one (for models with a mean structure, the factor mean of the reference group should also be fixed to 0). In addition, the factor loadings (as well as the intercepts for models with mean structures) of the RI(s) are constrained to be equal across all groups. In so doing, there is only one set of estimated coefficients that optimally reproduces the data. In other words, a multiple-group model is identified. Meanwhile, since other parameters are estimated in reference to the standardized factor in the reference group and selected RI(s), the scale of the 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 when an inappropriate item is chosen to be a RI, severe Type I or Type II errors are expected in testing factorial invariance; that is, truly invariant items could be detected erroneously as non-invariant items and vice versa (Johnson, Meade, & DuVernet, 2009; Yoon & Millsap, 2007). Selection of a RI determines whether the true status of invariance could be detected using the multiple-group CFA method. Methodologists have proposed a number of methods which allow researchers to select the proper RI (see Woods, 2009; Rivas, Stark, & Chernshenko, 2009; Meade & Wright, 2012; Shi et al. 2017; Tang, Shi, & Song, 2018). It is noted that other approaches were proposed, which allow researchers to test invariance without using any specific item as RI (e.g., Raykov, Marcoulides, & Millsap, 2013). For the proposed BSEM approach, we assume researchers could identify the multiple group model by selecting the proper RI.

7. We applied the cutoff suggested by Cheung et al. (2002) as it is the most cited criterion.

Table 1: Average Point Estimates and Coverage Rates for Bayesian 95% HDI

Mag	N	Factor Loadings			Intercepts		
		Population	Ave. Est.	CR	Population	Ave. Est.	CR
NA	100	**0.00**	-0.01	0.97	**0.00**	0.00	0.95
	200	**0.00**	-0.02	0.95	**0.00**	0.00	0.93
	500	**0.00**	0.00	0.96	**0.00**	0.00	0.96
	1,000	**0.00**	0.00	0.94	**0.00**	0.00	0.96
	2,000	**0.00**	0.00	0.96	**0.00**	0.00	0.95
	10,000	**0.00**	0.00	0.94	**0.00**	0.00	0.94
TR	100	**0.05**	0.04	0.97	**-0.05**	-0.05	0.95
	200	**0.05**	0.03	0.94	**-0.05**	-0.05	0.93
	500	**0.05**	0.05	0.96	**-0.05**	-0.05	0.96
	1,000	**0.05**	0.05	0.94	**-0.05**	-0.05	0.95
	2,000	**0.05**	0.05	0.96	**-0.05**	-0.05	0.95
	10,000	**0.05**	0.05	0.94	**-0.05**	-0.05	0.94
SM	100	**0.20**	0.19	0.96	**-0.30**	-0.29	0.95
	200	**0.20**	0.19	0.94	**-0.30**	-0.30	0.93
	500	**0.20**	0.20	0.96	**-0.30**	-0.30	0.96
	1,000	**0.20**	0.20	0.94	**-0.30**	-0.30	0.95
	2,000	**0.20**	0.20	0.96	**-0.30**	-0.30	0.95
	10,000	**0.20**	0.20	0.95	**-0.30**	-0.30	0.94
LG	100	**0.40**	0.39	0.96	**-0.60**	-0.58	0.95
	200	**0.40**	0.39	0.94	**-0.60**	-0.60	0.93
	500	**0.40**	0.40	0.95	**-0.60**	-0.60	0.96
	1,000	**0.40**	0.40	0.94	**-0.60**	-0.60	0.95
	2,000	**0.40**	0.40	0.96	**-0.60**	-0.60	0.95
	10,000	**0.40**	0.40	0.95	**-0.60**	-0.60	0.94

Note. Mag=Magnitude of non-invariance; NA=truly invariant; TR=trivial non-invariance; SM=small level of non-variance; LG=large level of non-invariance; Population= the population values of the difference parameter (D_{ij}); CR= coverage rate.

Table 2: Probabilities of Making Different Conclusions for Evaluating Invariance on Factor Loadings

Mag	N	BSEM with ROPE [-.10,.10]				BSEM with ROPE [-.20,.20]				Equivalence Test					LRT		ΔCFI	
		Prac. Invar.	Imp. Non.	Non.	Incon.	Prac. Invar.	Imp. Non.	Non.	Incon.	Exce.	Close	Fair	Medi.	Poor	Invar.	Non.	Invar.	Non.
	100	0.00	0.01	0.02	0.97	0.00	0.00	0.03	0.97	0.66	0.02	0.06	0.07	0.19	0.96	0.04	0.99	0.01
	200	0.00	0.00	0.05	0.95	0.11	0.00	0.05	0.84	0.64	0.06	0.12	0.07	0.11	0.94	0.06	1.00	0.00
NA	500	0.00	0.00	0.03	0.97	0.82	0.00	0.03	0.15	0.65	0.19	0.14	0.01	0.01	0.96	0.04	1.00	0.00
D=.00	1000	0.37	0.00	0.05	0.58	0.99	0.00	0.01	0.00	0.72	0.22	0.06	0.01	0.00	0.95	0.05	1.00	0.00
	2000	0.87	0.00	0.04	0.09	1.00	0.00	0.00	0.00	0.81	0.18	0.01	0.00	0.00	0.96	0.04	1.00	0.00
	10000	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.95	0.05	1.00	0.00
	100	0.00	0.01	0.04	0.96	0.00	0.00	0.04	0.96	0.63	0.02	0.06	0.05	0.25	0.92	0.08	0.99	0.01
	200	0.00	0.00	0.06	0.93	0.08	0.00	0.07	0.86	0.57	0.08	0.12	0.09	0.14	0.90	0.10	1.00	0.00
TR	500	0.00	0.00	0.13	0.86	0.72	0.00	0.14	0.14	0.51	0.21	0.14	0.09	0.04	0.85	0.15	1.00	0.00
D=.05	1000	0.16	0.00	0.22	0.62	0.93	0.00	0.07	0.00	0.40	0.32	0.19	0.06	0.02	0.76	0.24	1.00	0.00
	2000	0.42	0.00	0.36	0.22	0.99	0.00	0.01	0.00	0.33	0.49	0.16	0.01	0.01	0.61	0.39	1.00	0.00
	10000	0.96	0.00	0.04	0.00	1.00	0.00	0.00	0.00	0.89	0.09	0.00	0.02	0.00	0.03	0.97	1.00	0.00
	100	0.00	0.08	0.18	0.74	0.00	0.02	0.25	0.74	0.28	0.02	0.06	0.06	0.58	0.66	0.34	0.87	0.13
	200	0.00	0.14	0.32	0.54	0.00	0.02	0.44	0.54	0.16	0.03	0.10	0.09	0.62	0.48	0.52	0.92	0.08
SM	500	0.00	0.35	0.54	0.12	0.02	0.02	0.86	0.09	0.01	0.03	0.08	0.10	0.78	0.11	0.89	0.98	0.02
D=.20	1000	0.00	0.64	0.36	0.00	0.03	0.03	0.94	0.00	0.00	0.00	0.03	0.06	0.91	0.00	1.00	1.00	0.00
	2000	0.00	0.90	0.10	0.00	0.02	0.01	0.96	0.00	0.00	0.00	0.00	0.02	0.98	0.00	1.00	1.00	0.00
	10000	0.00	1.00	0.00	0.00	0.03	0.03	0.95	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	1.00	0.00
	100	0.00	0.51	0.24	0.25	0.00	0.28	0.47	0.25	0.04	0.00	0.01	0.01	0.94	0.17	0.83	0.43	0.57
	200	0.00	0.79	0.16	0.05	0.00	0.46	0.49	0.05	0.00	0.00	0.00	0.01	0.99	0.03	0.97	0.34	0.66
LG	500	0.00	1.00	0.00	0.00	0.00	0.87	0.13	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.15	0.85
D=.40	1000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.05	0.95
	2000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.01	0.99
	10000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00

Note. Mag=Magnitude of non-invariance; NA=truly invariant; TR=trivial non-invariance; SM=small level of non-variance; LG=large level of non-invariance; ROPE= region of piratical equivalence; Prac. Invar. = practically invariant; Imp. Non. = importantly non-invariant; Non.
Uncer. = non-invariant with uncertain practical importance; Incon.= inconclusive; LRT=likelihood ratio test.

Table 3: Probabilities of Making Different Conclusions for Evaluating Invariance on Intercepts

Mag	N	BSEM with ROPE [-.10,.10]				BSEM with ROPE [-.20,.20]				Equivalence Test					LRT		ΔCFI	
		Prac. Invar.	Imp. Non.	Non.	Incon.	Prac. Invar.	Imp. Non.	Non.	Incon.	Exce.	Close	Fair	Medi.	Poor	Invar.	Non.	Invar.	Non.
NA D=.00	100	0.00	0.00	0.05	0.95	0.00	0.00	0.05	0.95	0.68	0.03	0.05	0.05	0.18	0.94	0.06	0.99	0.01
	200	0.00	0.01	0.06	0.93	0.27	0.00	0.07	0.66	0.65	0.07	0.10	0.07	0.11	0.92	0.08	1.00	0.00
	500	0.00	0.00	0.04	0.96	0.93	0.00	0.04	0.03	0.65	0.20	0.13	0.02	0.00	0.96	0.04	1.00	0.00
	1000	0.49	0.00	0.04	0.46	1.00	0.00	0.00	0.00	0.67	0.28	0.04	0.01	0.00	0.96	0.04	1.00	0.00
	2000	0.92	0.00	0.04	0.04	1.00	0.00	0.00	0.00	0.80	0.19	0.00	0.00	0.00	0.95	0.05	1.00	0.00
	10000	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.95	0.05	1.00	0.00
TR D=.05	100	0.00	0.01	0.06	0.94	0.00	0.00	0.06	0.94	0.66	0.01	0.07	0.03	0.23	0.92	0.08	0.98	0.02
	200	0.00	0.01	0.08	0.90	0.21	0.00	0.10	0.69	0.54	0.08	0.14	0.08	0.15	0.89	0.11	0.99	0.01
	500	0.00	0.00	0.16	0.84	0.79	0.00	0.16	0.05	0.45	0.22	0.19	0.08	0.06	0.83	0.17	1.00	0.00
	1000	0.25	0.00	0.24	0.51	0.99	0.00	0.01	0.00	0.36	0.36	0.22	0.04	0.01	0.75	0.25	1.00	0.00
	2000	0.47	0.00	0.43	0.10	1.00	0.00	0.00	0.00	0.26	0.50	0.19	0.04	0.01	0.55	0.45	1.00	0.00
	10000	0.99	0.00	0.01	0.00	1.00	0.00	0.00	0.00	0.82	0.15	0.00	0.04	0.00	0.02	0.98	1.00	0.00
SM D=.30	100	0.00	0.29	0.36	0.35	0.00	0.10	0.55	0.35	0.06	0.01	0.02	0.02	0.89	0.29	0.71	0.65	0.35
	200	0.00	0.62	0.28	0.09	0.00	0.21	0.70	0.09	0.01	0.01	0.01	0.02	0.95	0.07	0.93	0.57	0.43
	500	0.00	0.97	0.03	0.00	0.00	0.46	0.54	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.51	0.49
	1000	0.00	1.00	0.00	0.00	0.00	0.71	0.29	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.47	0.53
	2000	0.00	1.00	0.00	0.00	0.00	0.97	0.03	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.41	0.59
	10000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.22	0.78
LG D=.60	100	0.00	0.97	0.02	0.01	0.00	0.88	0.11	0.01	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.04	0.96
	200	0.00	1.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.01	0.99
	500	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
	1000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
	2000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
	10000	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00

Note. Mag=Magnitude of non-invariance; NA=truly invariant; TR=trivial non-invariance; SM=small level of non-variance; LG=large level of non-invariance; ROPE= region of piratical equivalence; Prac. Invar. = practically invariant; Imp. Non. = importantly non-invariant; Non.
Uncer. = non-invariant with uncertain practical importance; Incon.= inconclusive; LRT=likelihood ratio test.

Table 4: The 15-Item CES-D Scale

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 5: Results for Invariance Testing using BSEM

Item #	$\hat{D}_{\lambda j}$ (95% HDI)		$\hat{D}_{\tau j}$ (95% HDI)	
Item 1	0.003	[-0.024,0.029]	0.036	[0.011,0.061]
Item 2	-0.016	[-0.042,0.009]	0.055	[0.030,0.079]
Item 3	0	-	0	-
Item 5	-0.050	[-0.075,-0.023]	-0.008	[-0.033,0.017]
Item 6	-0.031	[-0.059,-0.004]	-0.009	[-0.036,0.017]
Item 7	-0.056	[-0.083,-0.030]	-0.052	[-0.077,-0.026]
Item 9	-0.045	[-0.072,-0.018]	**-0.125**	**[-0.151,-0.100]**
Item 10	0.057	[0.031,0.084]	**0.147**	**[0.122,0.171]**
Item 11	-0.024	[-0.049,0.002]	**0.127**	**[0.102,0.151]**
Item 13	-0.063	[-0.090,-0.038]	**-0.159**	**[-0.184,-0.134]**
Item 14	-0.079*	[-0.106,-0.052]	-0.092*	[-0.117,-0.065]
Item 17	**0.206**	**[0.180,0.234]**	**0.298**	**[0.275,0.322]**
Item 18	0.053	[0.024,0.081]	0.091*	[0.065,0.116]
Item 19	-0.055	[-0.082,-0.029]	-0.098*	[-0.124,-0.072]
Item 20	0.042	[0.015,0.069]	-0.036	[-0.061,-0.012]

Note. HDI= highest density interval. The point estimates are based on the means from the posterior distributions. Item 3 was used as the reference indicator; importantly non-invariant parameters are **in bold**; asterisks (*) indicate non-invariant parameters with uncertain practical importance; the rest tested parameters are practically invariant.

Figure 1: Flowchart for Evaluating Factorial Invariance via Bayesian Parameter Estimation

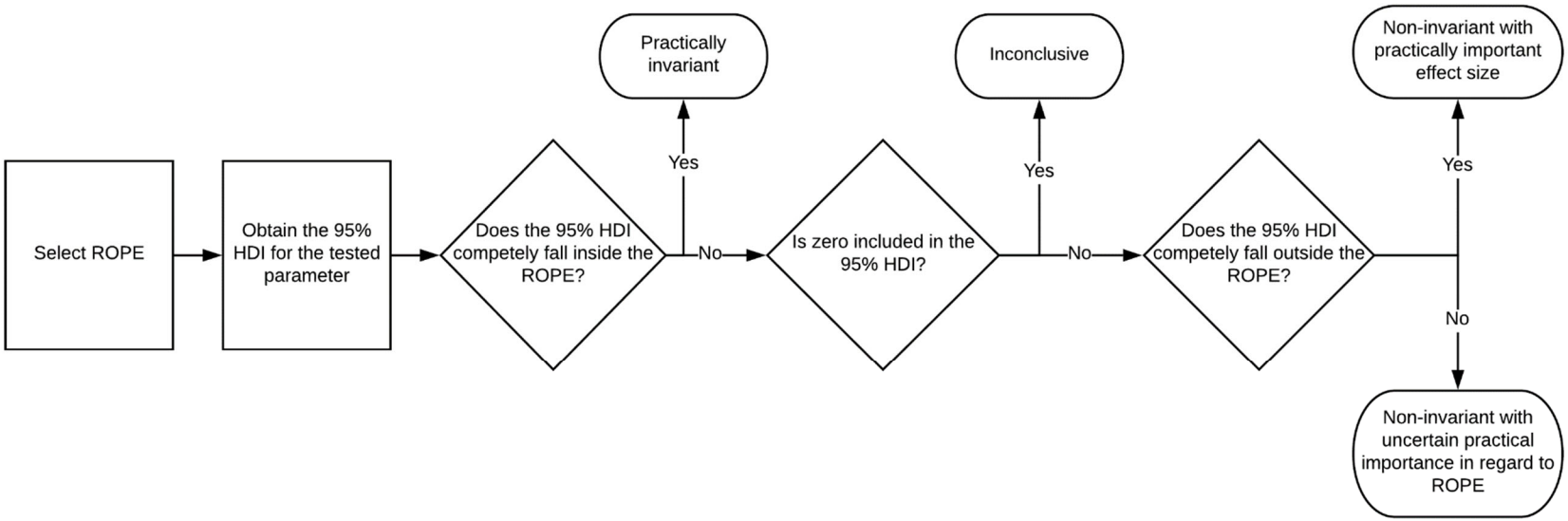


Figure 2: Trace plot of the standardized factor loading difference for item 1 ($D_{\lambda 1}$)

Figure 3: Trace plot of the standardized intercept difference for item 1 ($D_{\tau 1}$)

Figure 4: Posterior Distributions of the Cross-group Differences in the Tested Factor Loadings

Figure 5: Posterior Distributions of the Cross-group Differences in the Tested Intercepts

Notes for the Figures. HDI= highest density interval. ROPE= region of piratical equivalence

Figure 2:

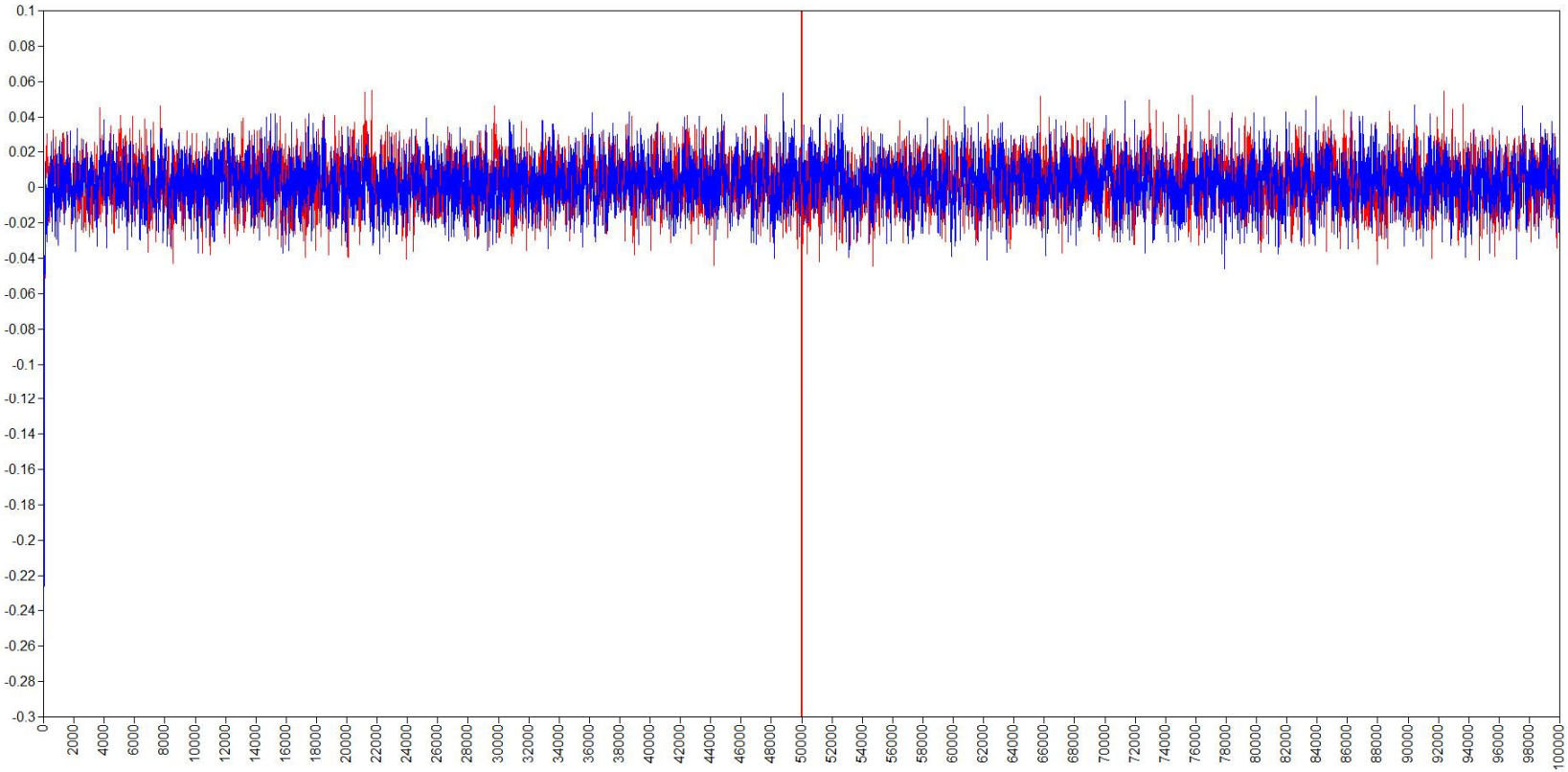

0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
-0.2
-0.22
-0.24
-0.26
-0.28
-0.3
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
50000
52000
54000
56000
58000
60000
62000
64000
66000
68000
70000
72000
74000
76000
78000
80000
82000
84000
86000
88000
90000
92000
94000
96000
98000
100000

Figure 3:

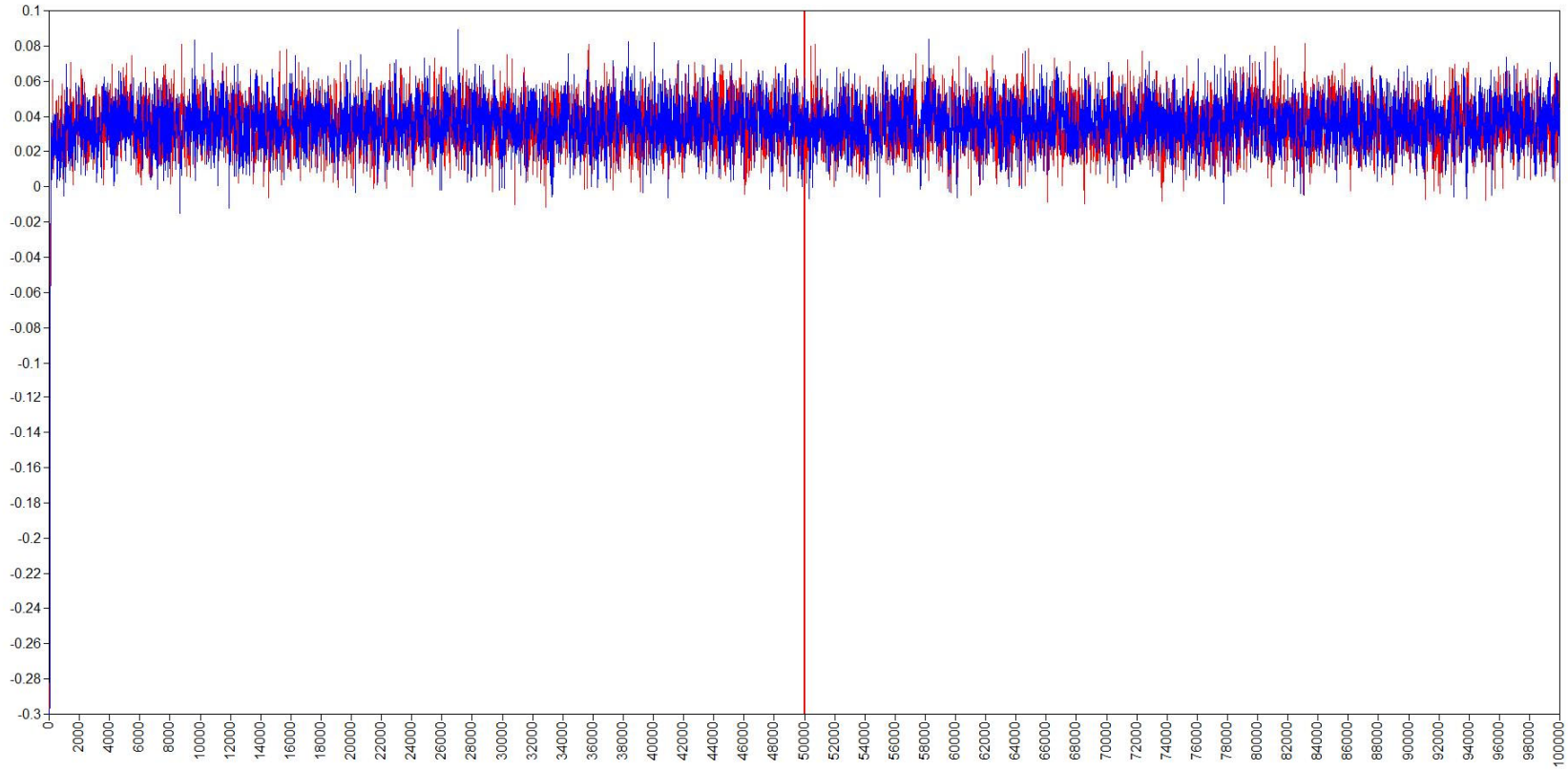

0.1
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.14
-0.16
-0.18
-0.2
-0.22
-0.24
-0.26
-0.28
-0.3
0
2000
4000
6000
8000
10000
12000
14000
16000
18000
20000
22000
24000
26000
28000
30000
32000
34000
36000
38000
40000
42000
44000
46000
48000
50000
52000
54000
56000
58000
60000
62000
64000
66000
68000
70000
72000
74000
76000
78000
80000
82000
84000
86000
88000
90000
92000
94000
96000
98000
100000

Figure 4:

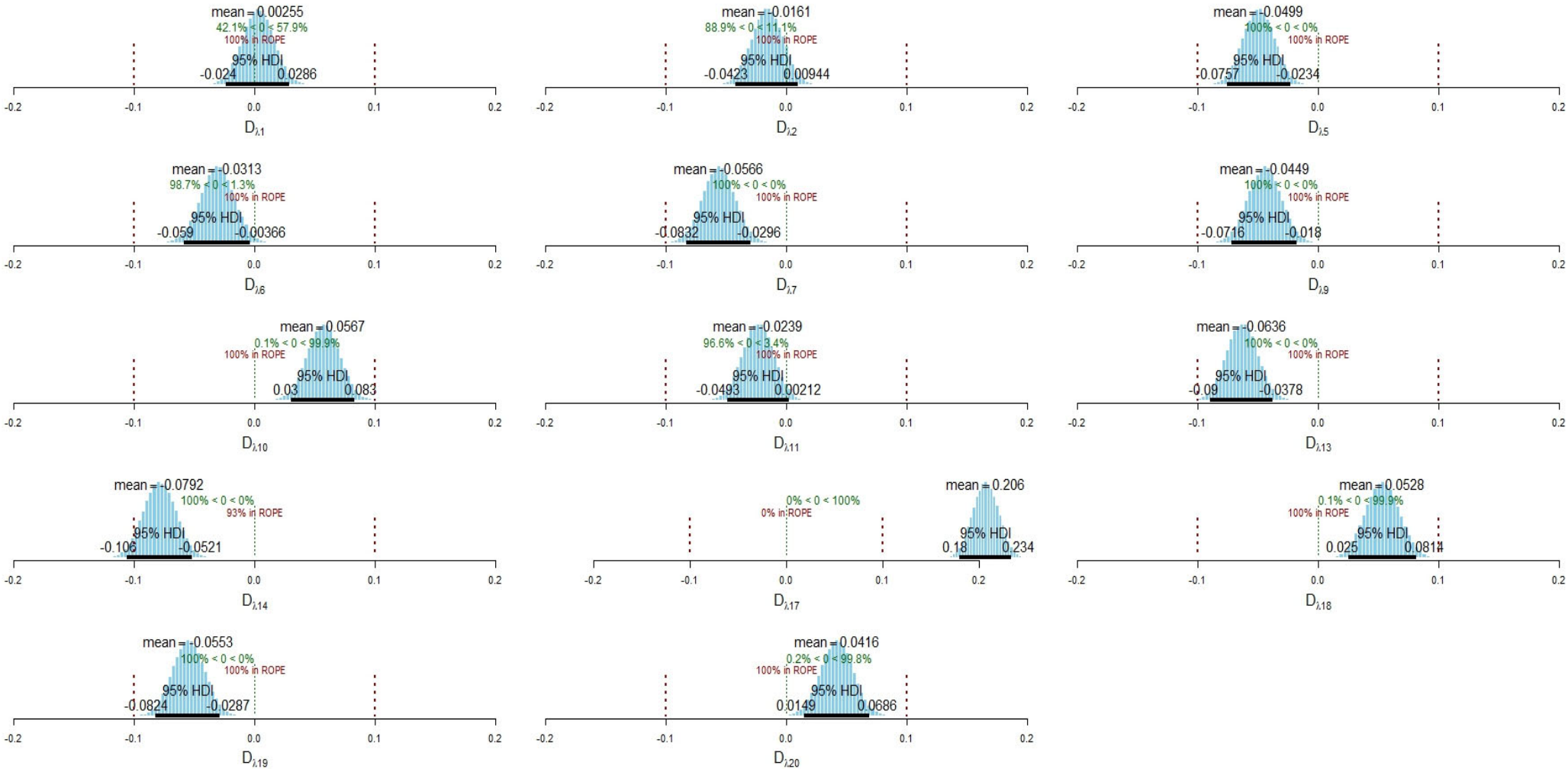
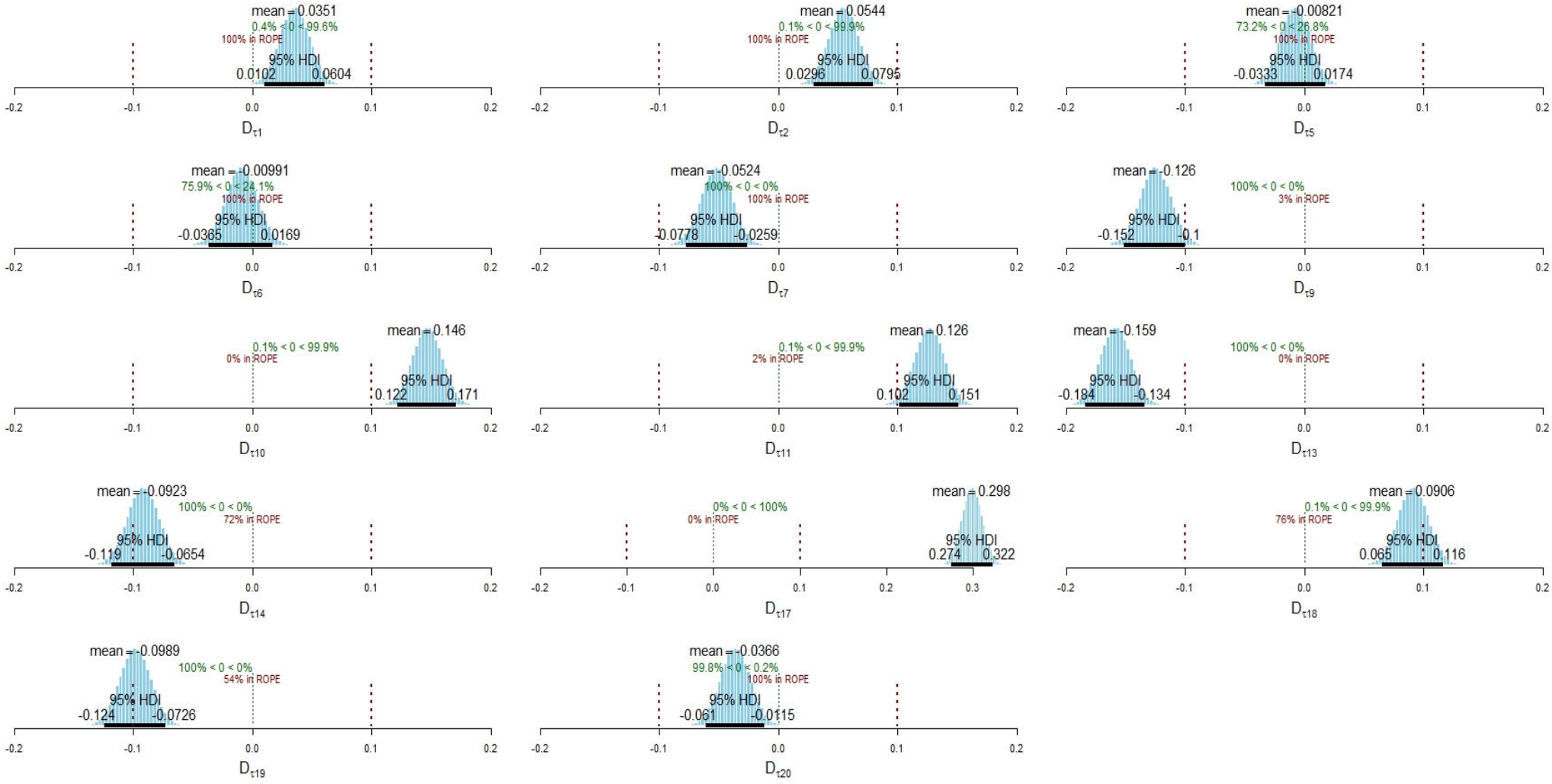

mean = 0.00255
42.1% < 0 < 57.9%
100% in ROPE
95% HDI
-0.024
0.0286
$D_{\lambda 1}$
mean = -0.0161
88.9% < 0 < 11.1%
100% in ROPE
95% HDI
-0.0423
0.00944
$D_{\lambda 2}$
mean = -0.0499
100% < 0 < 0%
100% in ROPE
95% HDI
-0.0757
-0.0234
$D_{\lambda 5}$
mean = -0.0313
98.7% < 0 < 1.3%
100% in ROPE
95% HDI
-0.059
-0.00366
$D_{\lambda 6}$
mean = -0.0566
100% < 0 < 0%
100% in ROPE
95% HDI
-0.0832
-0.0296
$D_{\lambda 7}$
mean = -0.0449
100% < 0 < 0%
100% in ROPE
95% HDI
-0.0716
-0.018
$D_{\lambda 9}$
mean = 0.0567
0.1% < 0 < 99.9%
100% in ROPE
95% HDI
0.03
0.083
$D_{\lambda 10}$
mean = -0.0239
96.6% < 0 < 3.4%
100% in ROPE
95% HDI
-0.0493
0.00212
$D_{\lambda 11}$
mean = -0.0636
100% < 0 < 0%
100% in ROPE
95% HDI
-0.09
-0.0378
$D_{\lambda 13}$
mean = -0.0792
100% < 0 < 0%
93% in ROPE
95% HDI
-0.106
-0.0521
$D_{\lambda 14}$
mean = 0.206
0% < 0 < 100%
0% in ROPE
95% HDI
0.18
0.234
$D_{\lambda 17}$
mean = 0.0528
0.1% < 0 < 99.9%
100% in ROPE
95% HDI
0.025
0.081
$D_{\lambda 18}$
mean = -0.0553
100% < 0 < 0%
100% in ROPE
95% HDI
-0.0824
-0.0287
$D_{\lambda 19}$
mean = 0.0416
0.2% < 0 < 99.8%
100% in ROPE
95% HDI
0.0149
0.0686
$D_{\lambda 20}$
-0.2
-0.1
0.0
0.1
0.2

Figure 5:


mean = 0.0351
0.4% < 0 < 99.6%
100% in ROPE
95% HDI
0.0102 0.0604
$D_{\tau 1}$
mean = 0.0544
0.1% < 0 < 99.9%
100% in ROPE
95% HDI
0.0296 0.0795
$D_{\tau 2}$
mean = -0.00821
73.2% < 0 < 26.8%
100% in ROPE
95% HDI
-0.0333 0.0174
$D_{\tau 5}$
mean = -0.00991
75.9% < 0 < 24.1%
100% in ROPE
95% HDI
-0.0365 0.0169
$D_{\tau 6}$
mean = -0.0524
100% < 0 < 0%
100% in ROPE
95% HDI
-0.0778 -0.0259
$D_{\tau 7}$
mean = -0.126
100% < 0 < 0%
3% in ROPE
95% HDI
-0.152 -0.1
$D_{\tau 9}$
mean = 0.146
0.1% < 0 < 99.9%
0% in ROPE
95% HDI
0.122 0.171
$D_{\tau 10}$
mean = 0.126
0.1% < 0 < 99.9%
2% in ROPE
95% HDI
0.102 0.151
$D_{\tau 11}$
mean = -0.159
100% < 0 < 0%
0% in ROPE
95% HDI
-0.184 -0.134
$D_{\tau 13}$
mean = -0.0923
100% < 0 < 0%
72% in ROPE
95% HDI
-0.119 -0.0654
$D_{\tau 14}$
mean = 0.298
0% < 0 < 100%
0% in ROPE
95% HDI
0.274 0.322
$D_{\tau 17}$
mean = 0.0906
0.1% < 0 < 99.9%
76% in ROPE
95% HDI
0.065 0.116
$D_{\tau 18}$
mean = -0.0989
100% < 0 < 0%
54% in ROPE
95% HDI
-0.124 -0.0726
$D_{\tau 19}$
mean = -0.0366
99.8% < 0 < 0.2%
100% in ROPE
95% HDI
-0.061 -0.0115
$D_{\tau 20}$

References

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. http://doi.org/10.1037//0033-2909.105.3.456

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. http://doi.org/10.1177/1745691611406920

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), 711–715. http://doi.org/10.1037/a0019917

French, B. F., & Finch, H. (2016). Factorial Invariance Testing under Different Levels of Partial Loading Invariance within a Multiple Group Confirmatory Factor Analysis Model. *Journal of Modern Applied Statistical Methods*, *15*(1), 511–538. http://doi.org/10.22237/jmasm/1462076700

French, B., & Finch, W. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*(3), 378–402. http://doi.org/10.1207/s15328007sem1303

Horn, J. L., & Mcardle, J. J. (1992). A Practical and Theoretical Guide to Measurement Invariance in Aging Research. *Experimental Aging Research*, *18*(3), 117–144. http://doi.org/10.1080/03610739208253916

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populatons. *Psychometrika*, *36*, 409–426.

Kaplan, D., & Depaoli, S. (2012). Bayesian Structural Equation Modeling. In *Handbook of*

Structural Equation Modeling (pp. 650–673).

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

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. http://doi.org/10.1177/0013164411427395

Kruschke, J. K., & Meredith, M. (2015). Package "BEST." Retrieved from https://cran.r-project.org/web/packages/BEST/BEST.pdf

Kruschke, J. K. (2011a). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, *6*(3), 299–312. http://doi.org/10.1177/1745691611406925

Kruschke, J. K. (2011b). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. http://doi.org/10.1037/a0029146

Kruschke, J. K., Aguinis, H., & Joo, H. (2012). The Time Has Come. *Organizational Research Methods*, *15*(4), 722–752. http://doi.org/10.1177/1094428112457829

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. http://doi.org/10.3758/s13423-016-1221-4

Lee, S.-Y. (2007). *Structural Equation Modeling: A Bayesian Approach*. Wiley.

Martin, A. (2008). Bayesian analysis. Retrieved from

https://deepblue.lib.umich.edu/bitstream/handle/2027.42/116259/Bayesian_Analysis.pdf?sequence=1&isAllowed=y

McGaw, B., & J??reskog, K. G. (1971). FACTORIAL INVARIANCE OF ABILITY MEASURES IN GROUPS DIFFERING IN INTELLIGENCE AND SOCIOECONOMIC STATUS. *British Journal of Mathematical and Statistical Psychology*, *24*(2), 154–168. http://doi.org/10.1111/j.2044-8317.1971.tb00463.x

Meade, A. W. (2010). A Taxonomy of Effect Size Measures for the Differential Functioning of Items and Scales. *Journal of Applied Psychology*, *95*(4), 728–743. http://doi.org/DOI: 10.1037/a0018966

Meade, A. W., & Lautenschlager, G. J. (2004). A Comparison of Item Response Theory and Confirmatory Factor Analytic Methodologies for Establishing Measurement Equivalence/Invariance. *Organizational Research Methods*, *7*(4), 361–388. http://doi.org/10.1177/1094428104268027

Meredith, W., & William Meredith. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. http://doi.org/10.1007/BF02294825

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

Muthén, B., & Asparouhov, T. (2012a). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. http://doi.org/10.1037/a0026802

Muthén, B., & Asparouhov, T. (2012b). New Developments in Mplus Version 7. Retrieved August 4, 2017, from http://www.statmodel.com/download/handouts/MuthenV7Part1.pdf

Muthén, L., & Muthén, B. (n.d.). BO 1998-2012. *Mplus User's Guide*.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence:

Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. http://doi.org/10.1037/a0022955

R Development Core Team. (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria.

Radloff, L. S. (1977). A Self-Report Depression Scale for Research in the General Population. *Appl. Psychol. Meas.*, *1*(3), 385–401. http://doi.org/10.1177/014662167700100306

Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, 1–17. http://doi.org/10.1177/1073191117711020

Shi, D., Song, H., Liao, X., Terry, R., & Snyder, L. A. (2017). Bayesian SEM for Specification Search Problems in Testing Factorial Invariance. *Multivariate Behavioral Research*, 1–15. http://doi.org/10.1080/00273171.2017.1306432

Song, X.-Y., & Lee, S.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: with applications in the medical and behavioral sciences*. http://doi.org/10.1002/9781118358887

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292–306. http://doi.org/10.1037/0021-9010.91.6.1292

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in

Cross-National Consumer Research. *Journal of Consumer Research*, *25*(1), 78–107. http://doi.org/10.1086/209528

Steiger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika*, *50*(3), 253–263.

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402–15. http://doi.org/10.1037/1082-989X.11.4.402

Tanner, M. A., & Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation: Rejoinder. *Journal of the American Statistical Association*, *82*(398), 548–550. http://doi.org/10.2307/2289463

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. http://doi.org/10.1177/109442810031002

Xie, Y., & Hu, J. (2014). An Introduction to the China Family Panel Studies (CFPS). *Chinese Sociological Review*, *47*(1), 3–29. http://doi.org/10.2753/CSA2162-0555470101