UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

BAYESIAN AND FREQUENTIST APPROACHES FOR FACTORIAL INVARIANCE TEST

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

YUTIAN TANG THOMPSON

Norman, Oklahoma
2018

BAYESIAN AND FREQUENTIST APPROACHES FOR FACTORIAL INVARIANCE TEST

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

Dr. Hairong Song, Chair

Dr. Robert Terry

Dr. Jorge Mendoza

Dr. Maeghan Hennessey

Dr. Lara Mayeux

For my love ones: Xiaoling Tang, Chan Wu and Cole Thompson

Acknowledgements

In choosing the subject as my dissertation, I purposely used “we” instead of “I”, because “we” represents the group of people who have supported me during my five-year Ph.D study program. This dissertation would not have been completed without them. Because of their love, intelligence and diligence, I was able to dedicate myself to my research and finally finish my dissertation. This is the perfect time to appreciate them and honor their help. First of all, I want to thank my advisor Dr. Hairong Song. She is a teacher, a friend, and a role model in my life. I appreciate her for bringing me into the world of quantitative psychology and helping me to become a researcher. In these past years, she shared with me her suggestions and advice, and gave me the freedom to pursue my research interests. With her whole-heartedly guidance, she helped made this dissertation what it is.

In addition, I have the pleasure in acknowledging my gratitude to my colleagues and scholars at University of Oklahoma. Dr. Dexin Shi, an excellent collaborator in sharing with me his advice on this work. I also wish to thank Dr. Robert Terry and Dr. Scott Wilson for giving me a job opportunity at the K20 center. It was such a wonderful experience to work with them. My committee members Dr. Jorge Mendoza, Dr. Lara Mayeux and Dr. Maeghan Hennessey provided their insightful suggestions and comments on my General Exam and dissertation.

Furthermore, I want to express my sincere appreciation to my family. I thank my parents Mr. Xiaoling Tang and Mrs. Chan Wu for always being with me, especially in tough times. Without their support I could not have finish my journey towards this degree. In addition, I would not forget to thank a person with my full heart. He is my

husband Mr. Cole Thompson, a great man who is practicing the meaning of unconditional love, has devoted himself completely to love and take care of his wife in these past five years.

Finally, I lift up my praise to God, the Almighty, my Lord, for His blessing and grace that provided me this great opportunity to study in this university with my group of friends and family.

Table of Contents

List of Tables

List of Figures

Abstract

Measurement invariance test concerns with whether the group membership is related to the attributes of a test. In the framework of Structural Equation Modeling (SEM), it is implemented by the multiple-group confirmatory factor analysis (CFA). Although this test has been widely applied in empirical studies, two research questions still need to be explored. One is how to appropriately choose the reference indicator (RI), the second is how to locate the non-invariant items. The challenge lies in appropriately choosing an invariant RI and accurately locating the non-invariant item parameters.

In this dissertation, we used two well-designed simulation studies to answer these two aforementioned questions. In Study I, we compared three commonly-used methods, named MaxL, Minχ^2 and BSEM and evaluated their performance of reference indicator selection. In Study II, we employed two Bayesian methods, Bayes factor and Bayesian estimation to locate the positions of non-invariance. All methods were applied to empirical datasets.

We found Minχ^2 and BSEM are superior to MaxL in correctly choosing the reference indicator. In addition we discovered Bayes factor can more accurately locate the non-invariant item parameters and distinguish the invariant items from the contaminated ones. Finally, the application of Cauchy prior will help to improve Bayes factor's performance.

Introduction

Measurement invariance (MI) is concerned with whether the group membership is related to the attributes of test. It often serves as an important tool in establishing measurement equivalence across groups, particularly when scores from self-report measures are being compared (Horn & McArdle, 1992; Meredith, 1993; Shi, Song, & Lewis, 2017). The test helps to examine to what degree observed differences reflect differences in the underlying, unobserved latent constructs across groups. Questions could be addressed with this technique, for instance, *does a mean difference in a measure of depression between males and females reflect entirely the gender difference in trait scores of depressions*? Or, *is the observed difference contaminated by differences in psychometrical properties of the measure across gender groups*?

In fact, if a measure indeed behaves differently across groups due to differences in social norms, cultural norms, or response tendencies, any comparison on the observed composites of this measure (such as *t*-test or ANOVA) will likely lead to ambiguous conclusion. Research has shown that departures from measurement equivalence weaken the accuracy of selection based on composite scores (Millsap & Kwok, 2004), and cross-group difference in composite scores could mostly reflect the difference in psychometrical properties of the measure in use (Steinmetz, 2011). Without testing for measurement invariance, one cannot be certain whether observed differences across groups truly indicate the underlying latent differences among constructs. Establishing measurement invariance has been increasingly recognized as a prerequisite for examining mean differences across groups or mean changes over time.

The approaches in both item response theory (IRT) and structural equation modeling (SEM) can handle the measurement invariance test. In the current dissertation, we mainly focus on the factorial invariance test under structural equation modeling (SEM) context. To test for measurement invariance, the factorial invariance tests are conducted using techniques of multiple-group confirmatory factor analysis (CFA; Byrne, Shavelson, & Muthén, 1989; Horn, McArdle, & Mason, 1983; Jöreskog, 1971; Meredith, 1993; Millsap, 2012; Steenkamp & Baumgartner, 1998; Widaman & Reise, 1997; also see Vandenberg & Lance, 2000 for a review). Based on the multiple group CFA model, there is a linear relationship defined between the observed scores X_i and its corresponding latent factor η_i as:

$$X_i = \tau_i + \lambda_i \eta_i + \varepsilon_i \qquad (1)$$

The observed scores X_i are for an item *I*. Item parameter include the intercept τ_i the factor loading λ_i and the residual term ε_i. Holding the presumptions that X_i multivariate-normally distributes, we use mean vector and variance and covariance matrix to describe X_i.

$$E(X_i) = \tau_i + \Lambda_i \kappa_i \qquad (2)$$

Where κ_i is an r*1 vector of factor means and Λ_i is a matrix of factor loadings.

$$Cov(X_i) = \Lambda_i \Phi_i \Lambda_i' + \Theta_i \qquad (3)$$

Where Φ_i is the covariance matrix for latent factors and Θ_i is a variance-covariance matrix among the residuals.

The procedures of factorial invariance are to test a series set of invariances on item parameters (τ_i, Λ_i, Θ_i). First, it starts with a baseline model, where the configuration of factorial structure is set to be identical across groups. All parameters are freely estimated in this model, except for those imposed with minimum constraints for model identification. Then a series of multiple-group CFA models are fitted through imposing an increasing number of equality constraints that correspond to increasing levels of invariance. For example, weak factorial invariance assumes all factor loadings Λ_i are numerically equivalent across groups. Strong factorial invariance assumes all intercepts τ_i, along with all factor loadings Λ_i are equal across groups (e.g., Widaman & Reise, 1997). The strictest invariance constraints all three item parameters (τ_i, Λ_i, Θ_i) to be equivalent. Tenability of specific equality constraints is determined by testing the significance of chi-square difference between the models with and without these constraints. For instance, in the test of metric invariance, it is often to compare the less restricted model (configural invariance) with a more restricted model in which only loadings constrained.

In testing for factorial invariance, a common method for identification is to constrain the factor loading (and intercept) of one particular item to be equal across groups. The item chosen for this purpose is referred to as a reference indicator (RI). All other parameters (except for the factor variance and mean for the selected group) are then freely estimated in reference to the scale of the chosen RI (Cheung & Rensvold, 1998; Johnson, Meade, & DuVernet, 2009; Meade & Wright, 2012).[1] However, as Rensvold and Cheung (1998, p.1022) pointed out, “This creates a dilemma. The reason one wishes to estimate the constrained model in the first place is to test for factorial invariance, yet the procedure requires a priori assumption of invariance with respect to the referents.” Whether the selected RI is truly invariant is critical in detecting invariance or non-invariance of other items. 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.

Despite of its importance, RI selection has still been under-addressed and inappropriately implemented in applied research. Using the keywords of *measurement invariance*, *measurement equivalence,* and *factorial invariance,* a recent search in the database of PsycINFO yielded a total of 192 applied studies published in 58 different peer-reviewed journals 2017. *Psychological Assessment*, *Developmental Psychology*, *PLOS one*, and *European Journal of Psychological Assessment* listed in order as the first four journals in terms of number of publications on factorial invariance related research. Surprisingly, only 13 of the reviewed studies (6.8%) mentioned RI selection. Ten of them selected “the first item” (whichever the first item was) as the reference indicator, and the other 3 did not state the specific method of their RI selection.

It is worth of noting that fixing factor variance of all groups to unity was found in 15 (7.8%) reviewed studies as the way for identification of multiple-group CFA models. This method could produce misleading results for factorial invariance tests (Rensvold & Cheung, 1998; Yoon & Millsap, 2007, Shi, Song, Liao, Terry, & Snyder, 2017), although it works well in identifying single-group models. Research has shown that if the imposed equality of factor variances does not hold in data, true differences in factor variances may be shifted to be observed differences in factor loadings across groups (Rensvold & Cheung, 1998). Results of factorial invariance tests would be invalid in this case. Therefore, this method is not recommended for identification of multiple groups in testing for factorial invariance (Yoon & Millsap, 2007).

Indeed, many different methods have been proposed on RI selection in recent years. Some originated from item response theory (IRT), and some are SEM-based approaches. However, there appears to be a large gap between methodological advances and empirical uses of RI selection in applied research. A need is called to further understand the imperativeness of RI selection and more importantly, to deeply understand the advantages and disadvantages of using different methods, providing useful guidance for future practices of factorial invariance test and its related analyses.

The goal of Study I is to meet this need by comprehensively evaluating and comparing a three selected, commonly-used methods of RI selection. They are named "$MaxL$", "$Min\chi^2$" and "Bayesian SEM". To this end, a simulation study was conducted in which a variety of data conditions were generated for multiple-group CFA models with continuous indicators. Power of correctly choosing a truly invariant item as RI serves as a major criterion for performance evaluation. Then, a large real-world data set of 12,811 respondents was used to empirically demonstrate and compare the uses of RI selection methods. Lastly, recommendations and suggestions were given based on the comparisons from both simulated and empirical investigations.

After selecting the invariant reference indicator properly, one can easily test the invariance of item parameters in terms of standard procedures of the multiple CFA model. However, the practice of factorial invariance on real data would not often be tenable at certain invariance levels. For example, the strong and strict levels seldomly hold invariance. Therefore, to correctly locate the non-invariance becomes highly necessary for further analysis on measurement equivalence. The "*location of non-invariance*" indicates the optimal separation of items with item-parameters differences from the ones that are invariant across-group. It indicates two aspects of meanings. One is to place the position of the non-invariant item; and the other one is to distinguish the invariant item from those contaminated ones.

The appropriate way to locate non-invariance will greatly benefit researchers. First, it would help them to properly conduct the partial measurement invariance test under empirical modeling settings. Fitting the multiple-CFA with certain free estimates can produce more accurate results than with the equally full parameter constraints (Shi, Song & Lewis, 2017; Muthén & Asparouhov, 2013). For example, when the measurement invariance is not held, fitting the second-order latent growth curve models with partial constraints yielded less biased estimates than the original full constraint model (Liao, X, 2012). Second, it would help to explore the potential causes to the non-invariance. For example, if a pair of factor loadings differs, the association between the item and latent construct might be stronger in one group than the other. Many reasons might result in this inequality, such as distinct understanding, translation barriers, cultural discrepancy and so on.

Largest modification index (MI) is one common method to locate the non-invariance in literatures (Yoon & Millsap, 2007). It starts with constraining all item parameters as a baseline model. When consulting with the largest value of MI (higher than the cut-off), this method frees only one constraint on parameters estimate in the focal group. This procedure then sequentially relaxes one constraint at a time until MI is no longer significant, or no larger than the cut-off 3.84, at $\alpha = 0.05$ (Yoon and Millsap, 2007). The MI value can be obtained from the likelihood ratio (LR) test on the nested models with one degree of freedom (MacCallum, Roznowski & Neocowitz, 1992). One can easily identify the non-invariant items which are freely estimated. This method is also named the *Sequential Max-mod*. It performs well under the conditions of fewer contaminated items, larger sample size and larger magnitude. However, it is also found to inflate Type I error due to the potential model misspecification in the baseline model (Yoon & Millsap, 2007; Kim & Yoon, 2011; Whittaker, 2012).

To address upon the high false positive rate, Jung and Yoon (2016) proposed another method named *Forward CI*. Opposite from the *largest MI* with full constraints on the baseline, it does not hold any constraints (except the reference indicator). Instead, the method creates a new parameter (E.g., $\lambda_g - \lambda_{g\prime}$) corresponding to the loadings or intercepts difference. Employing the maximum likelihood estimation, the confidence interval (CI) is to estimate the new parameters. If CI does not include zero, non-invariance is believed to be located on the tested parameter. Otherwise, it is believed to have no difference on item functioning. The simulation results indicated that the *Forward CI* is generally superior to the *Largest MI* with lower Type I and II error rates. Yet, this superiority is reduced when the cut off value of *Largest MI* adjusts to be more conservative from 3.84 to 6.65 (Jung and Yoon, 2016).

Both *Largest MI* and *Forward CI* are proposed on the framework of well used null hypothesis significance testing (NHST). The former one is heavily based on the *p* value in null hypothesis test. If the *p* value for the LR test is small enough, the Chi-square statistics difference between two nested models are significant and non-invariance exists. The latter one profoundly depends on the confidence intervals (CI) for model parameter estimates. If the range of CI does not contain zero, it is believed that non-invariance occurs. However, due to the natural limitations of NHST, both methods yield some inevitable defects. First, the non-significant results from LR test cannot be taken as the conclusive evidence for the invariant parameters. Regarding the null hypothesis protocol, one can identify the source of parameter non-invariance when $p <$ 0.05. Nevertheless, one cannot accept the null nor be sure the hold of invariance when p is > 0.05, because the interpretation of non-significant p value can be one of two possibilities. Either there is evidence to support the null (to "accept" null), or it is the lack of sufficient evidence that the data is insensitive to distinguish the theory from the null (nothing follows from the data) (Dienes, 2014). As long as NHST produces only one fixed estimate of parameter without any credibility about other parameters values, decisions are narrowed down to be dichotomous; either to reject or failed to reject null. (Brooks, 2003; Dienes, 2011; Kruschke, 2011; 2014; 2018).

What is more, the reliance of p value is a limited source of evaluation in hypothesis test. It would easily reveal the problems of large sample size fallacy (Lantz, 2013; Bergh, 2015). That is a statistically significant result may be meaningless from a study with a large sample size, because the actual difference is trivial and the effect size is small. This problem has been quite common in the applications of NHST to locate non-invariance. For example, simulations showed the Chi-square test of *Largest MI* is likely to reject the null in large sample size conditions, especially when the magnitude of non-invariance is small (Bentler & Bonett, 1980; Marsh, Hau & Grayson, 2005; Mead, 2010).

Comparing to the use of *p* value that decides whether or not a point parameter value would be rejected, the confidence interval in the method of *Forward CI* consists of a range in which the potential values of cross-group parameter differences might be covered. Unfortunately, in this range, it does not provide the probability for all the values (Kruschke & Liddell, 2018; Kruschke, 2014). We do not know how much more probable is the null value than other values. Therefore, even if CI contains zero, it does not necessarily mean that the parameter invariance equally holds. The probability of other values in CI might be higher than zero, indicating that there is still some amount of small non-invariance on parameters. Jung and Yoon's simulation results have shown this problem. When the magnitude of non-invariance is small and sample size is small, *Forward CI* (at 95%) has relatively large Type II error rates (Jung & Yoon, 2016).

Concerning the limitations of current methods, the Study II from the Bayesian perspective will introduce new methods which do not depend on the *p* value or CI to decide the location of non-invariance. Instead, we focus on the Bayesian hypothesis test executed by two categories of methods: *Bayesian estimation* (BE) and *Bayes Factor* (BF). Once we review each approach and their applications of measurement invariance, we will center our attention on the subsume methods in each category. For BE, we will elaborate on the method *region of practical equivalence* (ROPE) and its extended version *ROPE with zero* (ROPE_0). Savage-Dickey will be used to calculate BF. Furthermore, we will compare these Bayesian approaches under the context of testing the factorial invariance. Pros and cons will be listed and we will apply each method in real data as a pedagogical example for empirical users.

Study I: A Comparison on Methods of Reference Indicator Selection in Testing Factorial Invariance

Methods of RI Selection

Two major categories of approaches have been proposed to aid RI selection. One is all-others-as-anchors (AOAA) approach, and the other is Bayesian SEM (BSEM) approach. The AOAA approach originated from IRT, and has been considered as perhaps the only reasonable way to empirically identify RI while invariance status of all items is initially unknown. AOAA approach begins with fitting a baseline model in which all parameters are constrained to be equal across groups. Then each single item alternately serves as the target item, and parameters for the target item are to freely estimate while the others are still constrained to be equal. Then likelihood ratio (LR) test is used to compare the model fit between the two nested models, which is approximately χ^2 distributed with degrees of freedom equal to the difference in free parameters. The significance of this test indicates the presence of cross-group item differences.

The AOAA approach indeed subsumes two methods with different criteria for RI selection. The first one, labeled as *MaxL* in this study, chooses a RI as the item that produces non-significant LR statistics and meanwhile, has the largest factor loading (Stark, Chernyshenko, & Drasgow, 2006; Rivas, Stark, & Chernyshenko, 2009). This method has ever been recommended due to its high power of detecting item differences while controlling for nominal type I error (Meade & Wright, 2012). It could also outperform the BSEM approach in detecting item differences when large differences exist in factor loadings (Shi, Song, Liao, Terry, & Snyder, 2017). However, there is a methodological concern with this method. Woods (2009) stated that magnitude of factor loadings does not necessarily ensure item equivalence in using *MaxL* approach. For instance, when item *A* and item *B* both produce non-significant LR statistics, item *A* could be chosen as the RI due to its factor loading being the largest, even though item *B* is the one that indeed functions the same across groups but item *A*. In this case, *MaxL* would make a mistake in choosing a correct RI.

The second method, labeled as *Min*χ^2 in this study, selects a RI as the item that produces the smallest LR statistic among all items (Woods, 2009). The idea behind this approach is that the magnitude of LR statistic reflects the degree of difference in item functioning. So the smaller LR statistic is, the smaller the item difference is. This approach distinguishes itself from *MaxL* in that it does not require the smallest LR statistic to be non-significant. Woods (2009) showed that *Min*χ^2 performed well under a variety of data conditions in identifying truly invariant items with power rates of 90% and above.

The Bayesian SEM approach is a newly application of Bayesian method in testing for factorial invariance (Shi, Song, Liao, Terry, & Snyder, 2017; Shi, Song, Distefano, Maydeu-Olivares, McDaniel, & Jiang, 2018). It introduces a new parameter D_{ij} to represent a parameter difference across groups, which can index factor loading difference ($D_{loading}$) and intercept difference ($D_{intercept}$). A selection index for the *j*th item Δ_j can then be defined as a sum of standardized difference measures of $D_{loading}$ and $D_{intercept}$ for this item:

$$\Delta_j = \frac{|\widehat{D_{loading}}|}{SD_{loading}} + \frac{|\widehat{D_{intercept}}|}{SD_{intercept}} \tag{4}$$

where $\widehat{D_{loading}}$ and $\widehat{D_{intercept}}$ are respective estimates of difference in factor loadings and intercepts, and $SD_{loading}$ and $SD_{intercept}$ represent standard deviations of those differences.

The BSEM approach imposes informative priors with zero-mean and small-variance for $D_{loading}$ and $D_{intercept}$, which is referred to as "approximate identification constraints" (Muthen & Asparouhov, 2012). It ensures latent factors to be properly scaled and more importantly, makes $D_{loading}$ and $D_{intercept}$ estimable. Once $D_{loading}$ and $D_{intercept}$ are estimated for item *j*, one can compute the selection index Δ_j and then evaluate its posterior distribution. The item that produces the smallest posterior mean on Δ_j is considered to have the largest likelihood of being invariant across groups. This method produced high power of searching RI under a majority of simulation conditions (Shi, et al., 2017). It performed well especially when there were fewer non-invariant items with large magnitude of differences and large sample sizes. Power can be much higher than 0.90 when only 20% of items function differently across groups. The research showed that the choice of small prior variances did not significantly impact the power rates of RI selection.

Direction Effect and RI Selection

In previous research on RI selection, a two-group CFA model was typically used as the population model in data simulation. One group served as a reference group where factor means and variances were set to be known, and the other group served as a focal group where factor means and variance were freely estimated. A uniform direction of parameter differences was often simulated for simplicity. While factor loadings were simulated be the same for truly invariant items across groups, they were set to be smaller in focal group than those in reference group for items functioning differently (e.g., Stark, Chernyshenko & Drasgow, 2006; Woods, 2009; Meade & Wright, 2012; Shi, Song, Liao, Terry & Snyder, 2017). For instance, if the factor loadings were set to be .8, .8, .8, and .8 for all four items in the reference group, they were set to be .8, .6, .6, and .8 in the focal group. As a result, the truly invariant items (items 1 and 4 in the example) happened to have larger factor loadings than the non-invariant items (items 2 and 3 in the example). RI selection methods in favor of high loadings would have high power of selecting truly invariant items. However, such high power could just be the artifacts of data simulation with a uniform direction.

What if the direction of parameter differences is reversed? For instance, if the factor loadings are set to be .6, .6, .6, and .6 for all four items in the reference group, and .6, .8, .8, and .6 in the focal group, the methods in favor of high loadings are likely to choose either item 2 or item 3 as RI. In this case, the power of correctly selecting invariant items as RI would be low. Therefore, it is critical to consider the directions of parameter differences in generating data and evaluating power of the methods for RI selection.

In this study, we differentiated three types of directions of parameter differences. *Positive* direction refers to the case where parameter values are larger in focal group than reference group. *Negative* direction refers to the case where parameter values are smaller in focal group than reference group. The third direction is the *mixed* direction where certain parameters have in part larger and smaller values in one group than the other. If the power of RI selection is influenced by the directions of parameter differences, *direction effect* is said to occur.

As follows, we first presented a comprehensive simulation study, and then empirical applications of the three RI selection methods. A discussion will be given on theoretical and empirical issues of RI selection. Based upon the results of our study, we provided some guidelines on the uses of these methods for applied researchers. We also offered some suggestions on the simulation methodology for methodological researchers.

Monte Carlo Simulation Study

Data Conditions

The population model was a two-group CFA model with 10 items loaded on a single factor. One group served as reference group and the other served as focal group. The variables manipulated in the data simulation were listed as following:

Sample size: Continuous data were generated with *N* = 100, 200, 500 per group, representing small, medium, and large samples in typical psychological research. Both groups were simulated to have equal sizes in all conditions.

Location of difference: Item differences were simulated to occur on either factor loadings or intercepts, never on both at the same time.

Percentage of non-invariant items: Consistent with previous simulation research (e.g., French & Finch, 2008; Meade & Wright, 2012), we simulated data with either 20% or 40% of non-invariant items in this investigation. This corresponded to the cases where either 2 or 4 items (out of 10 items) function differently across the two groups.

Magnitude of difference: The magnitude of cross-group differences was set to 0.2 and 0.4 for factor loadings, and 0.3 and 0.6 for intercepts. The former values for the parameter differences were considered to be small, and the latter values were considered to be relatively large (Kim & Yoon, 2011; Kim, Yoon & Lee, 2012; Meade & Lautenschlager, 2004; Shi, Song, & Lewis, 2017).

Direction of cross-group difference: Three directions were manipulated for factor loadings and intercepts, including positive, negative, and mixed directions.

In total, 72 data conditions were generated by fully crossing 3 sample sizes, 2 locations of difference, 2 percentages of non-invariant items, 2 magnitudes of difference in parameters, and 3 directions of differences. Each condition had 500 replications.

Data Simulation

The factor mean and variance were set respectively to 0 and 1 in reference group. The raw factor loadings, intercepts, and unique variances were set to .8, 0, and .36, respectively, for all items. In focal groups, factor mean and variance were set to .5 and 1.2, respectively, and uniqueness were set to .36 for all items. All factor loadings and intercepts in focal groups were generated to be equal to those in reference groups, except for the items that were manipulated to be different under certain conditions. An example of two-group population CFA model is depicted in Figure 1.1 where 20% of factor loadings were set to be different with the negative direction in across-group differences.

Data Analysis

Three methods were used to analyze the simulated data, including *MaxL*, *Min*χ^2, and BSEM. In all analyses, the factor mean and variance were fixed to be 0 and 1, respectively in the reference groups. All the other parameters were freely estimated except for those required to be constrained by the procedures.

In using the *MaxL* method, the baseline model constrained all items to be equal across the focal and reference groups. Then, the equality constraints were relaxed for one item at time, yielding the reduced model. The differences in the target item were then examined using likelihood ratio test. This procedure was repeated for testing each of the other items in the model. Eventually, a reference indicator was chosen as the item that produced non-significant LR statistic and had the largest factor loading as well. When using the $Min\chi^2$ approach, the significance of LR statistic was not a concern; instead, the values of LR statistics were rank ordered for all items. A reference indicator was chosen as the item yielding the smallest LR.[2]

When using the BSEM method, the parameter D_{ij} was computed for each factor loading ($D_{loading}$) and each intercept ($D_{intercept}$) across groups. After imposing the normal priors of zero-mean and small-variance of 0.001 on the parameterD_{ij}, MCMC was run a minimum of 50,000 and maximum of 100,000 iterations. The estimates at every 10th iteration retained to form posterior distributions for factor loadings and intercepts. The means and standard deviations of these posterior distributions were then computed. Consequently, each item had a selection index Δ_j computed, indicating the summary of standardized difference in both factor loading and intercept. The item with the smallest value of Δ_j was selected as the reference indicator.

Results

We used power rates to evaluate the performance of each method. Power rate was calculated as the percentage of correctly identifying a truly invariant item as RI among 500 replications under each condition. In addition, ANOVAs were performed on power rates to test the main effects and interaction effects of all the six data variables.

The power rates under all data conditions were summarized in Table 1.1. An ANOVA was performed on these power rates to test the main effects of all the six data variables. As shown in Table 1.2, the main effects (see ANOVA 1) were not significant for *location*, *sample size*, and *magnitude of difference* (all *ps* > .05). However, the effect of *method* was significant ($F_{(2, 206)}$ = 25.507, *p* < .001), with *Min*χ^2 and BSEM performing better than *MaxL* (*p*s < .001). Figures 1.2 to 1.5 also showed that under multiple conditions, *MaxL* produced low power rates, and some of those were even lower than the power rates of selecting a random item as RI. This occurred in 50% of the conditions (12 of 24 in Table 1.1) when the direction of parameter differences was positive. However, this was not the case for *Min*χ^2 and BSEM. Neither of these two methods was associated with lower-than-random power rates.

The effect of *direction* was significant ($F_{(2, 206)}$ = 19.623, *p* < .001), and average power rate in positive condition was lower than that in negative and mixed conditions (*p*s < .001). The direction effect was evident. However, Figures 1.2 to 1.5 indicated that a) the direction effect was greater for *MaxL* than for *Min*χ^2 and BSEM, and b) factor loadings were more subjective to such direction effect than intercepts, suggesting the possibility of interaction among these data variables.

The effect of *percentage* was significant ($F_{(1, 206)}$ = 33.608, *p* < .001). Table 1.1 showed that 40% of items being different produced lower power rates than 20% of being different (*p* < .001). This occurred on factor loadings (as shown by differences between Figures 1.2 and 1.3) as well as on intercepts (as shown by differences in Figures 1.4 and 1.5).

Have examined the main effects, we now ran a full ANOVA model to include all main effects, two-way interactions, and three-way interactions among the six data variables. Our focus here was the significance of the interaction effects. In this model, four-way interactions cannot be examined due to the limitation of the dada; that is, there were very few scores in each cell without enough variation among them. Thus, this ANOVA was performed on power rates without four-way interactions being included. In total, there were 6 main effects, 15 two-way interactions, and 20 three-way interactions in this model. The results were presented as ANOVA 2 in Table 1.2. However only certain effects that bears direct importance were reported and interpreted in the following.

We first looked at the three-way interactions involving two-way interaction of *method* × *direction*. For a significant three-way interaction, we examined the two-way interaction at each level of the third variable. If a two-way interaction was significant at a certain level of the third variable, we then tested for simple effects of the data variables. Pairwise comparisons were made thereafter by using Bonferroni correction to adjust for the level of significance.

Table 1.2 showed that the following three-way interactions were significant: *method* × *direction* × *percentage* ($F_{(4,\ 110)} = 9.84$, $p < .001$), *method* × *direction* × *sample size* ($F_{(8,\ 110)} = 3.779$, $p < 0.001$), *method* × *direction* × *magnitude* ($F_{(4,\ 110)} = 7.964$, $p < .001$), and *method* × *direction* × *location* ($F_{(4,\ 110)} = 5.529$, $p < .001$). Then the two-way interaction of *method* × *direction* (Table 1.3) was significant at each level of *percentage* (20% and 40%), *sample size* (N = 100, 200, 500), *magnitude* (small and large), and *location* (loadings and intercepts). The interaction effects were displayed in

Figures S1.1-S1.4 in the in the supplementary appendix. As reported in Table 1.4, the subsequent pairwise comparisons showed that a) under *positive* condition, *Minχ*^{2} and BSEM consistently outperformed *MaxL* at all levels of *percentage, sample size, magnitude*, and *location*; b) however, this was true only for *percentage* = 40% and *magnitude* = *large* under *negative* condition; and c) under *mixed* condition the three methods did not performed differently.

We then examined the three-way interactions involving the two-way interaction of *method* × *magnitude*. Table 1.2 showed that all three-way interactions were significant: *method* × *magnitude* × *percentage* ($F_{(2,110)}$ = 9.400, *p* < .001), *method* × *magnitude* × *sample size* ($F_{(4,110)}$ = 7.642, *p* < .001), *method* × *magnitude* × *direction* ($F_{(4,110)}$ = 7.964, *p* < .001), and *method* × *magnitude* × *location* ($F_{(2,110)}$ = 7.056, *p* = 0.001). Figures S1.5 to S1.8 in the supplementary appendix display the two-way interactions of *method* × *magnitude* (Table 1.5) at each level of *percentage, sample size, direction*, and *location*. Table 1.6 showed the results from pairwise comparisons with Bonferroni correction for *p* values. When the between-group differences in parameters were small, *Minχ*^{2} and BSEM outperformed *MaxL* at *percentage* = 40%, *sample size* = 100, *direction = positive*, and *location = loadings*, and they did not perform differently under other conditions. When the parameter differences were large, *Minχ*^{2} and BSEM outperformed *MaxL* at *percentage* = 40%, *sample size* = 500, *direction = positive & negative*, and *location = intercepts*, and they did not perform differently under other conditions.

A Pedagogical Example

To demonstrate the empirical uses of the three methods, we applied all of them to select RIs using data from a large-scale project (12,811 participants) -- Psychological Wellbeing of Children of Rural-to-Urban Migrant Workers in China. The measurement chosen for this demonstration was from the Revised Child Anxiety and Depression Scale (RCADS, Chorpita, Yim, Moffitt, Umemoto & Francis, 2000). This self-report scale contains 47 items in total. However, only the items (18 items) related to generalized anxiety were used here for demonstration. Responses were scored on a Likert-scale of 1 to 4, corresponding to “Never”, “Sometimes”, “Quite Often”, and “Always”. The Cronbach’s α was 0.897 in this sample.

There were 7,356 male (57.4%) and 5,455 female (42.6%) child respondents in this sample. A two-group CFA was fitted to data, and *MaxL*, *Min*χ^2, and BSEM were used to find RIs. Eventually *MaxL* and *Min*χ^2 each produced 18 different values of LR statistics when comparing the baseline model and each reduced model. Then all 18 values were rank ordered from the smallest to largest. As shown in Table 1.7, item 7 in this scale was associated with the smallest LR statistic so that *Min*χ^2 chose this item as RI. For those items that yielded with non-significant LR statistic, item 7 was the one that had the largest factor loading in the baseline model. Thus *MaxL* chose Item 7 as the RI.

Then we used BSEM method to select a RI by specifying a two-group CFA model with the commands knownclass = c (g = 1 2) under Variable, and type = mixture; estimator = bayes; under Analysis (Muthen & Asparouhov, 2012). The parameter D_{ij}, representing a summarized difference of each item across groups, was set under model constraint. We imposed the normal prior of zero-mean and small variance (N (0, 0.001)) on each D_{ij} through the DIFF option under Model Priors. We let MCMC run for a minimum of 50,000 and a maximum of 100,000 iterations with thin = 10. The M*plus* output contained the necessary information for the posterior distribution of D_{ij} (including $D_{factor\_loading}$ and $D_{intercept}$). Table 1.8 showed the estimates for $D_{factor\_loading}$, $D_{intercept}$, and their standard deviations. The selection index Δ_j was then calculated using Equation 4 for each item. Eventually item 7 was chosen to be the RI because it produced the smallest Δ_j (= 0.646) out of 18 items.

Discussion

Inappropriate selection of reference indicators would jeopardize the outcome of factorial invariance test using multiple-group CFA approach. Unfortunately, the importance of RI selection has still not been fully aware among researchers (only 13 out of 198 reviewed articles mentioned something on RI selection). Meanwhile, pros and cons of current RI selection methods have not been well understood, which in part hinders the uses of these methods. In the present study, we aimed to address this issue by comparing a few commonly-used RI selection approaches, thereby providing certain guidelines on RI selection for applied researchers.

The simulation study revealed that *Min*χ^2 and BSEM performed better than *MaxL* in selecting correct item as reference indicator. This was particularly true under the *positive* condition where parameter values for functionally-different items were higher in the focal group than the reference group, regardless of the levels of all other conditions under investigation. Under the *negative* condition, *MaxL* performed much better than itself in the positive condition, and showed equivalent power as the other two under certain circumstances, such as small percentage of functionally-different items and small magnitude of cross-group difference in parameters. Under *mixed* condition, no significance differences were found for the three methods of being compared; however, *MaxL* appeared to be slightly inferior when the sample size and the loading difference were small.

The direction effect was evident in using *MaxL* approach. This was consistent with the expectation stated earlier in this article, that is, methods in favor of high loadings such as *MaxL* tend to perform poorly under conditions where truly invariant items happened to be the items with low factor loadings (i.e., positive condition). However, they would perform decently in most of cases when truly invariant items happened to be the items with high factor loadings (i.e., negative condition). This may in part explain why *MaxL* showed high power of correctly selecting RI in previous research where only negative condition was simulated (e.g., Meade & Wright, 2012). It appeared that non-uniformed direction of parameter differences (i.e., mixed condition) would remedy the drawback of favoring high loadings using *MaxL* approach. In this case, the power rates of detecting truly-invariant items were comparable among the three methods.

Another key feature of *MaxL* approach lies in the utility of LR statistic in testing for the significance of item difference between groups. Research has shown that the power of LR test is highly influenced by sample size and consequently, even very small difference in item parameters would lead to significant LR test when *N* is large (Ankenmann, Witt, & Dunbar, 1999; Meade, 2010). We found in our simulation analyses that when the percentage of functionally-different items was small, increasing sample size increased the power of detecting truly-invariant items. However, power decreased substantially or behaved inconsistently as sample size increased (to 500 for instance), particularly when both were large at the same time for the percentage of functionally-different items and the magnitude of item difference. This was true regardless whether the direction was positive or negative, and whether the difference occurred on factor loadings or intercepts. Thus high sensitivity to sample size makes *MaxL* approach not plausible to use in applied research.

Minχ^2 and BSEM approaches did not show any significant differences in their performance across all data conditions. However, when there were 40% of functionally-different items, the power rates of these two approaches were noticeably higher in negative condition than those in positive condition, which was only true for differences occurring in factor loadings. Our observation could be explained by the *reliability paradox* (see Hancock & Mueller, 2011). That is, when fitting SEM models, for a given level of model misspecification, better measurement quality is associated with poorer model fit (Heene, Hilbert, Draxler, & Ziegler, 2011; McNeish, An, & Hancock, 2018; Shi, Maydeu-Olivares, & Distefano, 2018; Shi, Lee, Maydeu-Olivares, 2018). In other words, the model misspecification (e.g. non-invariance) is “weighed” more heavily as the standardized factor loading becomes larger. It is also noted that the legitimacy of the *Min*χ^2 and BSEM approaches depends on certain assumptions; namely, the latent variables in multiple-group CFA model should be scaled in the way that the metric of the model parameters can be considered as a good approximation to the metric otherwise set by truly invariant parameter(s) only (see Shi et al., 2017). Therefore, the ideal condition for the *Min*χ^2 and BSEM approaches is when the majority of the tested items are invariant, and/or the non-invariant items are given lighter weights (i.e., with smaller factor loadings). Under the *positive* condition, the non-invariant items were simulated to have larger standardized factor loadings (than the truly invariant items); thus, given that the non-invariant items are more heavily “weighed”, the power of selecting the proper RI is expected to be suboptimal, especially when the number of non-invariant items is large (e.g., 40%). Future studies are needed to explore the role of the measurement quality (i.e., the size of the standardized factor loadings) on the accuracy of RI selection.

The properly choosing RI will let the item parameters be estimable in reference to the scale of an invariant item. Then, the standard procedures of factorial invariance test can be executed by a series set of invariant constraints. However, in the real empirical applications, the invariance of each item parameter are more often failed to be tenable. The normal procedures of factorial invariance can show whether the non-invariance exists, according to the significant results from likelihood ratio test. Unfortunately, it is not able to specify the exact positions of those contaminated parameters. Therefore, the methods to locate the non-invariance become highly necessary. In the following Study II, we employed three up-to-day methods from Bayesian hypothesis test, trying to answer the research question: how to locate the non-invariance?

Study II: An Investigation of Bayesian Analysis in Factorial Invariance Test

Introduction of Bayesian Analysis for Hypothesis Test

Previous literatures have discussed the difference between Bayesian and traditional NHST in performing hypothesis tests. (Brooks, 2003; Bayarri, M. J., & Berger, 2004; Dienes, 2011; Gelman, Carlin, Stern, Dunson, Vehtari & Rubin, 2013; Stegmueller, 2013; Kruschke, 2011; Kruschke, 2014; Kruschke & Liddell, 2018) The primary difference is the parameter is treated as a fixed constant for NHST but as a random variable for Bayesian approach. NHST aims at the probability of getting only the best-fitting parameter value, although such a parameter value depends heavily on the *p* value assuming the null is true. In contrast, Bayesian analysis focuses on the probabilities of all candidate parameter values. Given the observed data, it updates the prior to the posterior distribution of the credibility over all possible parameter values.

More specifically, suppose a model *m* with unknown candidates of parameters *θ* given by data *D*, the Bayesian theorem produces the posterior distribution of *θ*. It is

$$p(\theta_m \,\big|\, D, m) = \frac{p_m(D|\theta_m,m)p_m(\theta_m|m)}{\sum_m \int d\theta_m p_m(D|\theta_m,m)p_m(\theta_m|m)} \qquad (5)$$

The equation can also be extended as:

$$p(\theta_1, \theta_2, \ldots, m \,\big|\, D) = \frac{p(D|\theta_1,\theta_2,\ldots,m)p(\theta_1,\theta_2,\ldots,m)}{\sum_m \int d\theta_m p(D|\theta_1,\theta_2,\ldots,m)p(\theta_1,\theta_2,\ldots,m)} \qquad (6)$$

The numerator $p(D|\theta_1, \theta_2, \ldots, m)$ is the likelihood function for the data conditioning on the parameters of models and $p(\theta_1, \theta_2, \ldots, m)$ is the prior probability of parameter *θ*. The nominator is the integration representing the average of marginal likelihood $p_m(D|\theta_m, m)$ based on the model across all values of *θ*, weighted by the prior probability of θ. After taking the data into account, the posterior distribution $p(\theta_m \mid D, m)$ therefore has been updated from the prior state of belief on parameters by the likelihood function.

Nevertheless, to properly obtain the posterior distribution $p(\theta_m \mid D, m)$ is typically difficult by the traditional numerical integration. As the number of model parameters increases, the high dimensional parameter space involves the combinations of all possible parameter values. It requires assessing the likelihood function for each combination of parameter values and letting them combine with the prior to derive the posterior analytically (Van Ravenzwaaij, Cassey and Brown, 2018). However, no such computation is available in practice. Fortunately, this problem has been solved by the application of Markov Chain Monte Carlo (MCMC). It is a computer-based sampling method which repeatedly draws the random samples from the posterior distribution and summarizes the statistics of each draw. MCMC greatly benefits users particularly in Bayesian inference, as it approximates the property of posterior distributions.

MCMC procedure begins with an initial sample from the distribution, and then generates a proposal sample with some added random noise. Based on the plausibility, MCMC then needs to decide to accept or reject the newly proposed sample. If the proposal draw has a higher posterior value than the initial sample, MCMC accepts the proposal as the new sample for the next iteration. If the proposal draw is not higher, then it is designed to either accept or reject the sample by random chance. If the proposal is rejected, MCMC only needs to copy the initial sample use it for the next iteration. It will repeatedly run this procedure until enough samples are available.

Previous studies have described this process of MCMC as the *Metropolis algorithm* (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1953).

Although Metropolis algorithm has been useful in practice, it is less efficient for the proposal distribution which is usually too broad or too narrow (Kruschke, 2014). In other words, Metropolis algorithm will tend to reject the proposal sample frequently when parameters are strongly correlated. Consequently, in order to get the right posterior, the algorithm must run continuously with a larger number of samples. To improve its efficiency, a new method called Gibbs sampling was introduced, (Geman & Geman,1987, Gelfand & Smith, 1990; Smith & Roberts, 1993) which follows most of the same steps of Metropolis except drawing samples from the parameters' conditional distributions (Van Ravenzwaaij, Cassey & Brown, 2018). Each sample would not be drawn randomly, but instead from the probability distribution of parameter that depends on the value of another parameter. Therefore, it improves the efficiency to generate the posterior distribution. In this study, we used Gibbs sampling to run the MCMC chain.

Keeping the general Bayesian rules in mind, we will introduce its applications in Bayesian hypothesis test. There are two main approaches recommended in recent literature, Bayesian estimation (BE) and Bayes factor (BF).

Bayesian Estimation

Bayesian estimation focuses on the space of all possible parameter values. Taking the data into account, it starts with updating the belief (prior) on each parameter value to a posterior distribution based on Bayes' rules. Then, by using the posterior distribution, it makes the inferential statements for the parameter θ of interest (Rouder, Haaf and Vandekerckhove, 2018).

The *highest density interval* (HDI) can be used for the parameters inference. It provides a range of highly credible values for parameter θ. The estimate point inside the interval has a higher credibility than those outside. As a summary statistic, 95% of probable parameter values of HDI is often used to test the null (or other interested value). For example, one can simply reject the null value that falls outside a posterior 95% HDI. However, this method is not able to determine whether the null value should be accepted or withdraw. To address upon this problem, Kruschke (2011, 2014, 2018) proposed a new decision rule which employs a small range of parameter values around the null called the *region of practical equivalence* (ROPE). The values within this range are taken as equivalent as null value. If the entire ROPE lies outside the 95% HDI of posterior distribution of parameters, one can reject null. If the entire ROPE completely contains 95% HDI, contrast to the traditional NHST, one can truly "*accept*" null. However, if ROPE and HDI partially overlap, ROPE cannot completely cover 95% HDI, and no more concrete decisions can be made. The presented data are insufficient to make any decisions between reject or accept null, but "*uncertain*" for the hypothesis testing. Therefore, the size of ROPE matters. A wide ROPE range will increase the probability to accept the null and Type II error. Yet, a too narrow ROPE might be more likely overlap the HDI, increasing the uncertainty rate. There are no standard rules to specify the size of ROPE, because the range of ROPE highly depends on its practical purposes (Serlin & Lapsley, 1993; Kruschke, 2014). For example, to test measurement invariance, the set of ROPE range is indeed associated with how trivial the error in which the invariance can be defined. The range of [-0.1, 0.1] indicates that cross-group parameters are still consider to be invariance, even if they have 0.1 amount of functioning difference.

Shi, Song, Distefano, Maydeu-Olivares, McDaniel & Jiang (2018) provided a new logic to improve the practice of ROPE. Holding the same decisions when ROPE either completely includes 95% HDI or completely not, they extended two more situations as ROPE partially overlaps with 95% HDI. The first situation is when the point of zero is included within the 95% HDI, but is still inconclusive on the tested parameters. The second situation zero is excluded and is conclusive to reject null. Back to the example of measurement invariance, when 95% HDI does not contain zero, even if it partially overlaps with ROPE interval, they take the parameters to be non-invariance. Since this method is related with the point of zero, we named it ROPE with zero (ROPE_0).

The approach of Bayesian estimation is insensitive to the choice of prior distribution (Rouder, Haaf and Vandekerckhove, 2018) because when its incorporated with data it can gain the sufficient information and "overwhelms" the initial belief on parameters. In addition, because Bayesian estimation plays as a role to compromise between prior information and the data, therefore posterior distribution is heavily impacted to a greater extent by the data if the sample size is large (Gelman, et al., 2013).

Bayes Factor

Bayes factor (BF) was initially proposed by Sir Harold Jeffreys (1935, 1961) who contributed in the field of Bayesian hypothesis testing. The method was designed to test the null from the perspective of the Bayesian model comparison. It compares the probability of the data between two models, in which one model sets the parameter to zero (can also be other interested values), as the null hypothesis (H_0). The other model allows all possible parameters that are not equal to zero as the alternative hypothesis (H_1). More specifically, given the data, *D*, we have two models m = 1 and m = 2. The likelihood function *p*(D|m = 1) and *p*(D|m = 2), the priors are *p*(m = 1) and *p*(m = 2) for model 1 and model 2 respectively. Following the Bayes' rules in Equation 5, the posterior probability for model 1 and 2 can be taken as:

$$p(\theta_{m=1} \big| D, m = 1) = \frac{p_{m=1}(D|\theta_{m=1}, m=1) p_{m=1}(\theta_{m=1}|m=1)}{\sum_m \int d\theta_m p_m(D|\theta_m, m) p_m(\theta_m|m)} \quad (7)$$

$$p(\theta_{m=2} \big| D, m = 2) = \frac{p_{m=2}(D|\theta_{m=2}, m=2) p_{m=2}(\theta_{m=2}|m=2)}{\sum_m \int d\theta_m p_m(D|\theta_m, m) p_m(\theta_m|m)} \quad (8)$$

Let Equation 7 be divided by Equation 8.

$$\frac{p(\theta_{m=1} \big| D, m=1)}{p(\theta_{m=2} \big| D, m=2)} = \frac{p_{m=1}(D|\theta_{m=1}, m = 1) p_{m=1}(\theta_{m=1}|m = 1) / \sum_m \int d\theta_m p_m(D|\theta_m, m) p_m(\theta_m|m)}{p_{m=2}(D|\theta_{m=2}, m = 2) p_{m=2}(\theta_{m=2}|m = 2) / \sum_m \int d\theta_m p_m(D|\theta_m, m) p_m(\theta_m|m)}$$

The denominator of Equation 7 and 8 are the integrations of the likelihood function weighted by the prior over the all possible parameter values within hypothesis (Myung, 2003; Ly, Verhagen and Wagenmakers, 2016). The ratio of the two integrations is equal to 1, because the space of potential parameter values from both models are expected to be the same. Therefore, the following Equation 9 has three components:

$$\frac{p(\theta_{m=1} \big| D, m=1)}{p(\theta_{m=2} \big| D, m=2)} = \frac{p_{m=1}(D|\theta_{m=1}, m = 1)}{p_{m=2}(D|\theta_{m=2}, m = 2)} \times \frac{p_{m=1}(\theta_{m=1}|m = 1)}{p_{m=2}(\theta_{m=2}|m = 2)}$$

(9)

- *Prior odds:* $\frac{p_{m=1}(\theta_{m=1}|m = 1)}{p_{m=2}(\theta_{m=2}|m = 2)}$ represents the researchers' initial belief on each hypothesis before the data is given.
- *Posterior odds*: $\frac{p(\theta_{m=1} \big| D, m=1)}{p(\theta_{m=2} \big| D, m=2)}$ quantifies the relative plausibility of two models after receiving data.

- Bayes Factor: $\frac{p_{m=1}(D|\theta_{m=1}, m = 1)}{p_{m=2}(D|\theta_{m=2}, m = 2)}$ indicates how much change would be from the "*prior odds*" to "*posterior odds*".

BF is straightforward for the hypothesis test. When $BF_{01} > 3$, it is three times likely for the data under H_0 than H_1 and accept the null. While $BF_{01} < 1/3$, the data is three times more likely under H_1 than H_0, as the evidence to reject null^3. If BF_{01} is between 1/3 and 3, it is uncertain for any decision. Some literatures also suggest to consider a strong cut-off, say $BF_{01} > 10$ strongly supports for H_0 (Jeffreys, 1961; Kass and Raftery, 1995).

As a new alternative to NHST and *p* value, Bayes factor has been increasingly used not only in the forms of various applications in psychology area (Matzke, Nieuwenhuis, Rijn, Slagter, Molen, and Wagenmakers, 2015; Van Den Hout, Gangemi, Mancini, Engelhard, Rijkeboer, Dams, and Klugkist, 2014, 2017; Wong, & Schoot, 2012; Kary, Taylor, & Donkin, 2016; Mou, Berteletti, & Hyde, 2018) but also in the tutorial for common practices (Klugkist, Wesel & Bullens, 2011; Hoijtink, Béland, & Vermeulen, 2014; Mulder & Wagenmakers, 2016; Van De Schoot, Zondervan-Zwijnenburg and Depaoli, 2017). In addition, psychological applications are particularly in support of BF as it is especially suitable for testing a point null hypothesis (Williams, Bååth & Philipp, 2017). Imagine the model H_0 for null hypothesis gathers the probability mass exactly at point of zero, while H_1 holds the remainder of the probability spreading out across the range of the alternative values. It would be easy to compare the two hypotheses in terms of the ratio of their marginal likelihood (BF).

BF has several exceptional advantages for empirical practice. First, it is able to evaluate the information when the data is in favor of accepting null. If the result of p value is non-significant in the classical NHST, it fails to reject the null (but does not mean it can equally accept null). The *p* value is simple criteria which tends to overestimate the evidence against null but lacks the evidence for null (Mulder & Wagenmakers, 2016). However, results of BF can provide the evidence for acceptance of null hypothesis. The following literatures will provide a better understanding about BF's superiority on the aspect of accepting null in hypothesis testing. Bem (2011) conducted nine experiments to demonstrate the existence of psi in which future events effect on people's responses. In addition, Wagenmakers, Wetzels, Borsboom and Maas (2011) reanalyzed the data using Bayesian *t*-test. One of Bem's experiments tested the retroactive induction of boredom on neutral stimuli. They hypothesized that the test subjects who are high in stimulus seeking would also be significantly decreasing their liking for the target. However, their results indicated that $t(199) = -1.31$, $p = 0.096$, $d = 0.09$ which failed to reject null, but still no evidence to accept the null. Nevertheless, Wagenmakers et al. later used BF to substantially support the null, as BF_{01} was 7.6. Therefore, those test subjects high in stimulus seeking showed no difference in liking for the target from those who were not.

Second, Bayes factor is able to provide the information of uncertainty. Recall NHST cannot distinguish the "non-significant" results of the null hypothesis, either accepted or withdrawn for uncertainty. In comparison, BF does not have this issue. Several cut-off values defined the BF values clearly into three decision categories: reject, accept, or uncertain the null. To get a better sense of this feature, let us review some additional published articles that used NHST and BF respectively. Gollwitzer and Melzer (2012) tested the “Macbeth effect” which indicated the desire for people to cleanse themselves physically (called “moral cleansing”) when their moral selves have been threatened. The test subjects included both experienced and inexperienced participants to play one of two violent video games. One, in which, involved the violence against other humans and the other one was against an object. After the game, they were asked to pick up gifts in which half of them were hygiene products. Researchers accounted the number of choosing hygiene products as the measurement to test “Moral cleansing”. By applying ANOVA, the results indicated the inexperienced player chose more hygiene products after playing the violent game against humans rather than the objects $t(34) = -2.03$, $p = 0.05$, $d = 0.68$. Yet, no significant results ($\alpha = 0.05$) were found for experienced players $t(32) = 1.49$, $p = 0.15$, $d = 0.51$. However, Konijin, Schoot, Winter and Ferguson (2015) used Bayes factor to re-analyze the data. The results of BF about experienced player was BF = 0.87. It was an anecdotal evidence for the alternative, according to Jeffreys’ classification scheme of BF (Jefreys, 1961). The data was only 0.87 times as likely to have occurred under H_1, leaving some uncertainty and researchers cannot make any concrete decisions. Such ambiguity does not literally mean the results are unclear, but instead gives us a more profoundly insight of the relationship between the magnitude of uncertainty and statistical decisions.

Like other statistical methods, Bayes factor has its own fallacies. Most critics complain about the cut-off values (Dienes, 2014; Gigerenzer & Marewski, 2015; Kruschke & Liddell, 2018). Since BF is incapable of providing the direct evidence for probabilities of hypothesis, it needs some form of criteria to make decisions instead.

Jeffreys (1961) suggested 3 (1/3), 10 (1/10), and 100 (1/100) to divide the values of BF into several categories of hypothesis decisions. Unfortunately, these arbitrary cut-offs might easily allow BF slip away and back to the suffering of old tendency to interpret *p* value. It may only give us another new looking but old rough tools only for dichotomous yes or no decisions. For example, Konijn et al (2015) is concerned about the results of BF are too similar to the hacking-behaviors which *p*-values possess. That is BF 3.01 is considered as substantial evidence, while 2.99 becomes anecdotal evidence. In fact, "*God would love a Bayes factor of 3.01 nearly as much as BF of 2.99*" (Rosnow and Rosenthal, 1989). Therefore, researchers have suggested being cautious on BF's interpretations (Konijn et al, 2015).

In addition, Bayes factor has been also under the criticism about its severe sensitivity to the choice of priors (Myung & Pitt, 1997; Kruschke, 2011; Kruschke, 2014; Kruschke & Liddell, 2018). BF essentially is the ratio of marginal likelihood. Unlike Bayesian estimation in which the data can provide sufficient information to overwhelm the impact from priors, marginal likelihood is highly sensitive to the prior distribution (Liu and Aitkin, 2008). For example, when the prior is able to provide more probability mass around the place where the likelihood distribution peaks, the marginal likelihood will increase. Yet, if the prior comes up with little probability mass on likelihood distribution, the marginal likelihood will be small (Kruschke, 2014). Liu et al. (2008) performed a simulation study which found the bias of BF heavily depends on the prior distribution for H_1. They used informative priors which express specific and defined information on parameters. They also used non-informative priors in which the distributions are diffused in a broad range. The results revealed the differences of priors impacting on BF_{10}. Comparing to the non-informative priors (Uniform, Jeffreys and Haldane), BF is highly biased in favor of H_1 when it is with an informative prior.

The Bayesian Applications in Measurement Invariance Test

In recent years, more applications of Bayesian approaches have been in the research area of measurement invariance (MI) test. For example, Shi et al. (2017) used the Bayesian Structural Equation Modeling (BSEM) under the multiple-group CFA model to locate the non-invariance. They introduced a new parameter D_{ij}, representing a parameter difference. It can index both factor loading difference ($D_{loading}$) and intercept difference ($D_{intercept}$). A selection index (in Equation 4) for the j^{th} item Δ_j can then be defined as a sum of standardized difference measures of $D_{loading}$ and $D_{intercept}$ for this item:

$$\Delta_j = \frac{|\widehat{D_{loading}}|}{SD_{loading}} + \frac{|\widehat{D_{intercept}}|}{SD_{intercept}}$$

where $\widehat{D_{loading}}$ and $\widehat{D_{intercept}}$ are respective estimates of difference in factor loadings and intercepts, and $SD_{loading}$ and $SD_{intercept}$ are standard deviations of those differences. By imposing the informative priors with zero-mean and small-variance, the method let $D_{loading}$ and $D_{intercept}$ of each item to be estimable. The invariance tests are carried out by 95% HDI across all D_{ij}, given the selected reference indicator. If HDI for D_{ij} fails to contain zero, the corresponding item parameters are not equal across groups. This method is a great application of informative and small variance priors to locate non-invariance. In this chapter, we will pay more attention on the studies which focus on the applications of Bayesian hypothesis test for MI.

Dr. Verhagen and her colleagues (2016) were the first to introduce Bayes factor to MI through the multiple group IRT models. To test invariance by null hypothesis, they took the difference between item parameter across groups to be zero as the null (a point, H_0: $d_j = 0$, for all j items) while the rest of all possible non-zero values were for the alternative hypothesis (an area, H_1: $d_j \neq 0$). BF is the ratio of the marginal likelihoods for the results of two hypotheses.

$$BF_{01} = \frac{p_{H_0}\left(D \middle| \theta_{H_0}, H_0\right)}{p_{H_1}\left(D \middle| \theta_{H_1}, H_1\right)} = \frac{p_{H_0}\left(D \middle| d_j = 0\right)}{\int p_{H_1}\left(D \middle| d_j \neq 0\right) p_1(d_j) d d_j}$$

(10)

where $p_1(d_j)$ is the prior distribution for alternative hypothesis. Instead of doing the integration for the marginal likelihood of all plausible values for alternative hypothesis weighted by priors, they practiced the Savage-Dickey density ratio (Dickey, J. M., & Lientz, 1970; Dickey, 1971; Wagenmakers, Lodewyckx, Kuriyal and Grasman, 2010):

$$BF_{01} = \frac{p_{H_0}\left(D \middle| \theta_{H_0}, H_0\right)}{p_{H_1}\left(D \middle| \theta_{H_1}, H_1\right)} = \frac{P\left(d_j = 0 \middle| H_1,\ D\right)}{P\left(d_j = 0 \middle| H_1\right)}$$

(11)

The standard computation of BF asks for the analytical integration out of all possible model parameter for H_1. Compared to that, the calculation of Savage-Dickey is simple. At the point of interest, BF only considers H_1 when dividing the height of the posterior by the height of the prior for parameters (Wagenmakers, et al, 2010). To apply Savage-Dickey in MI, the parameter invariance should be tested simultaneously within the MCMC sampling scheme. BF is the probability distribution of null hypothesis under the posterior $P\left(d_j = 0 \middle| H_1,\ D\right)$ divide the prior $P\left(d_j = 0 \middle| H_1\right)$ under the alternative H_1. For more details about the mathematical calculation of the Savage-Dickey ratio, please read Wagenmakers, et al (2010).

Verhagen et al (2016) used both multivariate normal and multivariate Cauchy priors in their simulation. The BF_{01} value equal to three was taken as a cut-off. Fifty replications were generated for each condition which included three different sample sizes plus two priors. Non-invariance was generated on five out of ten items difficulties, with the magnitude from 0, 0.1, 0.3 to 0.7. The results of their simulation indicated that conditions in which BF with Cauchy prior generally perform better than Normal prior. For example, about 91% to 97% of the invariant items with Cauchy were successfully identified as invariance. From 78% to 91%, items with Normal priors were able to be identified. In addition, the power rate of BF locating non-invariance was higher when the magnitude was large. BF, at 95% and above, accurately figured out the 0.7 amount of difference when sample size increased to more than N = 500 each group. However, the rate dropped down even less than 60% for the same conditions, when the magnitude of non-invariance decreased to 0.1. Finally, the rate of uncertainty (no concrete evidence) followed the same patterns across conditions. The point in which uncertainty reached the highest peak was when the non-invariance magnitude varied between 0.3 and 0.5. Neither small nor large amount of parameter differences could increase the rate of uncertainty.

Several advantages should be greatly emphasized as BF is applied in MI. First, it provides the convincing evidence to distinguish the decisions of truly invariant from uncertainty, letting researchers be more comfortable choosing invariant items. What is more, the item parameters when BF is between 1/3 and 3 do not literately mean the unsureness to make any decisions. Instead, it provides a general picture on how likely the parameter in focal group might deviate from the reference group. For instance, if BF_{01} is 1.5, it means the observed data is still in favor of H_0 and 1.5 times more likely than H_1. What it implies is the corresponding item might be possible to maintain the invariance if more information can be handed over. Second, BF is able to provide acceptable good power rates, particularly when the magnitude of parameter difference increases to the extent on upper-middle level. For instance, the simulation compared the BF with Wald test (for more details read Langer, 2008; Woods, Cai and Mang, 2012), a common method in Frequentist to test the non-invariance under IRT framework. The results showed neither BF nor Wald were superior to the other, when the magnitude of parameter difference was either 0.3 below or 0.5 above. However, when the difference was between 0.3 and 0.5, BF had relatively higher power rate than Wald test with critical level at 0.01[6].

However, there are some limitations of this study we should not ignore. First, as an alternative method, Bayesian estimation had not been taken into account. Only Bayes factor was applied in this study. What is more, the usage of BF was limited under IRT context. No more applications for testing factorial invariance under SEM framework. Second, Verhagen et al (2016) merely focused on the invariance of item difficulties. They did not design the test on item discriminations. Less information was provided about BF's performance both on item difficulties and discriminations. In addition, they used an insufficient number of replications in Monte Carlo simulation. With only 50 replications, it might be possible to produce biased estimates, since some particular samples might be more likely to arise than others (Bandalos, 1997). Finally, the study had no options for Uniform prior, a special form of ignorance about the true rate and assigns the prior probability equally on each possible count (Liu and Aitkin, 2008). On one hand, the density of Uniform is low around zero. It is therefore expected to convey less information about parameter difference than Cauchy. On the other hand, since its distribution spreads out within a certain range, the previous simulation indicated BF was slightly more in favor of H_1 over H_0, comparing to other non-informative priors (Liu and Aitkin, 2008).

In the current study, we introduce both Bayes factor and Bayesian estimation to locate the non-invariance under SEM framework. Our main purpose is to provide a more comprehensive Bayesian perspective to locate non-invariance. The methods will show common users how they function under the multiple-CFA models for factorial invariance test. We specifically show how Bayesian approaches make decisions on accepting the invariance, detecting the uncertainty, and locating the non-invariance on item parameters.

Monte Carlo Simulation Study

The current simulation study went through three steps. First, we generated data into different conditions according to the study design. Then, we applied both Bayes factor and Bayesian estimation on the generated data and tested null hypothesis. The methods produced three decisions: accept null (the parameter is invariant across groups), deny null (non-invariance exists), or show no evidence to conclude. Finally, based on the decisions, we evaluated the results by calculating the power rate, uncertainty rate and rate of correctly locate invariant items (rate of invariance).

Data Conditions

We used a two-group CFA population model to generate the multivariate normal data. Ten items loaded on a single latent factor for each group. One group served as the reference group, and the other one served as the focal group. The variables manipulated in the simulation were listed as following conditions:

Sample Size: Continuous data were generated with balanced *N* = 100, 200, 500 each group and unbalanced *N* = 250 and *N* = 500 for reference and focal group respectively. The sample size increased from 100 to 500, representing small, medium and large samples in typical psychology research.

Non-invarianct items: Eight out of ten items (80%) were generated with non-invariance. The non-invariant variable started from the second item in the model. The first and the last item kept the invariance. The magnitudes of rest 8 items increased in the order of 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7 (Figure 2.1).

Location of difference: Item differences were simulated to occur on either factor loadings or intercepts, never both at the same time.

Prior: there were 3 different non-informative prior distributions used: Uniform, Cauchy and Normal distributions. The Uniform prior (range between -100 and 100) is the distribution that equally assigns probability to each of the counts (Liu and Aitkin, 2008). The normal density distributes with a mean zero and a variance of two. And the Cauchy (0, 1) is transformed as a *t* distribution with 1 degree of freedom, for the easy execution in JAGS.

Therefore, 192 data conditions were totally generated by fully crossing 4 sample sizes, 2 locations of difference, 8 magnitudes of difference on parameters and 3 prior distributions.

Data Simulation

The factor mean and variance were generated to 0 and 1 respectively in the reference group. The raw factor loadings, intercepts and unique variance were set to 0.8, 0 and 0.36 accordingly for all items. All factor loadings and intercepts in focal groups were generated to be equal to those in reference groups, except for the items that were manipulated to be different under certain conditions. However, the 8 magnitudes of non-invariance were generated starting from the second to the ninth item simultaneously (Figure 2.1). Therefore, 80% of observed variables obtained the differences between groups. There were total 500 replications in each condition. Mplus 7.1 was used to generate the data, given its higher operation speed. JAGS 4.3 was practiced under R-3.4.2 to analyze and summarize the results. We also applied other R packages to complete this study. They mainly included: “R2jags”, “runjags”, “MBESS”, “MCMCpack”, “logspline”, “HDInterval” et al.

Data Analysis

Both Bayes factor and Bayesian estimation (ROPE and ROPE_0) were used to analyze the simulated data. In the analysis, we fixed the factor mean and variance to be 0 and 1 only for the reference group. The 10th variable was taken as the reference indicator for model identification. The 10th factor loading, and intercept therefore were fixed to be same across groups. The rest of other parameters were freely estimated simultaneously.

For each factor loading and each intercept, the parameter D_{ij} was computed as the parameter difference between groups. Applying the non-informative prior distribution, each of three MCMC chains ran with 10,000 iterations after 500 burins.

The posterior distribution for each estimated parameter were finally constructed by these 10,000 draws. By examining the posterior distribution of parameterD_{ij}, both Bayes factor and Bayesian estimation were used to evaluate the invariance. Specifically, for Bayes factor, we used Savage-Dickey density ratio to get the BF_{01} value (Dickey, 1971). If BF_{01} was equal or larger than 3, we accepted the null, believing invariance had been held. If BF_{01} was equal or less than 0.33, we rejected null and non-invariance existed on the item parameter. If BF_{01} was between 0.33 and 3, we failed to make any concrete decisions. For Bayesian estimation, 95% HDI can be obtained from the posterior distribution. We set the ROPE limits between -0.1 and 0.1. There were two ways to make decisions. The first used the ROPE limits only. When the ROPE fully contained the 95% HDI, we accepted null. When ROPE was completely outside the 95% HDI, we rejected null. When ROPE had partially overlapped with 95% HDI, it was uncertain to make any decisions. The second one was to use both [-0.1, 0.1] limits and the point of zero. When 95% HDI completely fell into ROPE, we accepted the null and the invariance had been held, but if not, non-invariance was on parameter. However, if 95% HDI partially overlapped with ROPE, we followed the new logic (Shi, et al., 2018). 95% HDI contained zero, it was inconclusive. If not, we still believed the non-invariance existed though it was with the uncertain practical importance to some extent.

We used three criteria as the index to evaluate each method on parameter (non)invariance: the power rate, the rate to identify invariant item, and the uncertainty rate. The power rate was calculated as the percentage of accurately reject null (for non-invariant parameters) among 500 replications under each condition. The power computation of method *ROPE with zero* is based on two parts. One part was the rate of

95% HDI completely outside range of ROPE and the other part of rate was 95% HDI partially overlap with ROPE but exclude zero. We summed up two parts together as the final power. For the rate to correctly identify invariant item, it was the percentage of accurately accepting the null (for invariant parameters) among 500 replications. Since the first item was generated to be the same across groups, the rate would be calculated only from this item. For the uncertainty rate, it was the percentage of uncertain decisions among 500 replications under each condition. We expected the higher the power rate and the higher rate to identify invariant but lower uncertainty rate, the better the method could locate the non-invariance or detect the invariance.

To further investigate the main effects and interaction effects of all conditions on three criteria, we performed the Bayesian ANOVAs using the software package JASP (Wagenmakers, Love, Marsman, Jamil, Ly, Verhagen et al, 2018). It ran with default Cauchy prior (0, 0.5), in which the distribution centered in 0 with interquartile range $r = 0.5$. We also used R package “ggplot2” to visualize the results.

Results

Power Rate of Bayes Factor

Table 2.1 and Table 2.2 summarize the original rate of BF to reject and accept null hypothesis. The first column of Table 2.2 represents the rate to successfully identify invariant parameters. From the second to the ninth columns of Table 2.1, it shows the power to correctly locate non-invariance. The method BF represents several features. First, its power rate increases when the magnitude of non-invariance expands. In addition, the rate improves more rapidly on loadings than on intercepts, which generally leads the power of loadings (*M* = 0.609, *SD* = 0.366) to be higher than intercepts (*M* = 0.442, *SD* = 0.411). For example, the power of BF with Cauchy prior on loadings reaches to 0.75, when non-invariance difference is 0.3. On intercepts, the power approaches 0.75 until magnitude extends up to 0.5.

What is more, Figure 2.2 shows Bayes factor is highly sensitive to the choices of priors. The power rates of three priors vary from one another. Among them, the Cauchy presents the best, but the Uniform condition displays the worst on each level of non-invariance magnitudes. For instance, the power of Uniform prior on average is 0.446 across conditions, while it is 0.576 and 0.555 for Cauchy and normal prior respectively. Further investigation by Bayesian two-way ANOVA provides us more information in Table 2.3[5]. The first column named “Models” lists five models: the “Null model” only has the grand mean without any predictors. The second row “prior” model add only one predictor “prior” in the model. It is the same for the third row model with single “magnitude”. The forth model contains both prior and magnitude main effect alone. The final full model, not only keeps both the main effect but also interaction effect between prior and magnitude. Column “P(M)” is prior model probabilities, which has been set to be equal across all models. Column “P(M|data)” indicates the posterior model probability given by updated observed data. The next “BF_M” column is the most useful because it shows the change from prior to posterior model odds. The larger the value, the more likely the data supports the model by increasing the model credibility. For example, BF_M yields the highest value 35.709 for the model with two main effects on loadings, indicating this model with two main effect *priors* and *magnitude* receives the support from data. As evidence, the main effect of *prior difference* clearly represents BF’s sensitivity. The following “BF_{10}” column provides the Bayes factor of each row model against the first row null model. It is 1.714e +33 (we mark it “>10.000”) in favor of the two main effects model, with 0.986 in the final “% error” column. This error is similar to the coefficient variation in frequentist analysis. It provides the size of error in the integration relative to the Bayes factor (Wagenmakers, Love, Marsman, Jamil, Ly, Verhagen et al, 2018).

Furthermore, we notice that power rates are positively associated with sample sizes. The larger the sample size, the higher the power along with magnitude increasing (Figure 2.3). Generally, along with magnitude increasing, the power of larger sample size rises up much more quickly and stays in the higher level than small sample size conditions. For instance, the average power of sample *N* = 500 is 0.678. It is much higher than the average power (0.329) of a sample of only one hundred. What is more, we note this association is consistent across three prior conditions. However, the distinctions among three prior conditions needs further attention. First, the choices of Cauchy and Normal prior are considerably superior to Uniform prior, because the power of both prior conditions are much higher than Uniform condition on each level of sample size. For example, holding on the same 0.3 amount of parameter difference, the average power in larger sample size (*N* = 500) conditions is 0.677 with Uniform prior, but it is 0.942 with Cauchy and 0.936 with Normal priors. Similarly, for the same 0.3 non-invariance with small sample size (*N* = 100), the power mean is 0.119 with Uniform, yet it is 0.275 with Cauchy and 0.219 with Normal prior. Second, there is an interaction effect between *magnitude* and *sample size* in the Uniform condition. The power of the large sample is not improving as much with smaller samples, when magnitudes of differences increase. However, the interactions effects are absent in the other two priors conditions (Table 2.4). Be noticed, Figure 2.3 also supports that power of unbalanced sample is better than larger sample (N = 500) to some certain extent.

Power Rate of Bayesian Estimation

Table 2.5 and Table 2.6 provide the original rate of ROPE to reject and accept null hypothesis. Table 2.7 and Table 2.8 are about the rate of method ROPE_0. Figure 2.4 and Figure 2.5 display the power of two methods Bayesian estimation. Similar to BF, both ROPE and ROPE_0 increase the power when the magnitude expands. Unlike BF, power on loadings increases much slower and stays at a lower level than on intercepts. In addition, being consistent with the previous literatures, both methods are insensitive to the choice of priors. They exhibit a very similar power rate for each prior condition. Although the Uniform prior conditions obtain slighter higher power rate on intercept, further investigation indicates that it is not statistically superior to other two priors (Table 2.9). The data is not in support of any models with the effect *prior* for both non-invariance placing on loadings and intercepts. For instance, only BF_M of single main effect *magnitude* model obtains the largest values ($BF_{M\_ROPE\_intercept}$ = 9.360; $BF_{M\_ROPE\_0\_loading}$ = 9.550, indicating the data increases its probability mostly on these models. The "BF_{10}" column provide the Bayes factor of each row model against the first row null model. ALL BF_{10} of models for main effect *prior* are smaller than the cut-off 1/3, meaning that data is highly in favor of null model than the alternative. In other words, the power rate of *prior* conditions does not differ from each other.

Both methods share the similar patterns of BF that larger the sample size, higher the power rate (Figure 2.6 & Figure 2.7). Further analysis of Bayesian ANOVA shows more features of two methods (Table 2.10). First, the main effect of sample size has not been supported by the data. For instance, BF_{10_ROPE} is 0.162 for Uniform condition, representing this model obtains only 0.162 times more likely to the alternative model than null, given the observed data. Second, data are in favor of different models across prior conditions. For Uniform, the model with single magnitude main effect has been supported (BF_{M_ROPE} = 15.514, $BF_{M\_ROPE\_0}$ = 15.968), while the model with two main effects is instead preferred for Cauchy (BF_{M_ROPE} = 7.425, $BF_{M\_ROPE\_0}$ = 5.271) and Normal priors (BF_{M_ROPE} = 7.173, $BF_{M\_ROPE\_0}$ = 5.385). Third, the data does not support the model with interaction effect, but the corresponding BF_{10} values are mostly larger than the cut-off value 3. Though this model has its faults, it is still quite different from the null model.

Uncertainty of Bayes Factor and Bayesian Estimation

Uncertainty rate (Table 2.11 to Table 2.13) is another important criterion to determine the performance of methods. The lower the uncertainty, the better the method performs. First, Figure 2.8 demonstrates the uncertainty of method BF, showing the uncertainty has been controlled well in a low level (less than 0.4), no matter where the non-invariance locates. A bell curve appears roughly along with the horizontal magnitude scales. Moreover, a clear interaction effect between *priors* and *magnitude* exist on loadings. The rate of Uniform is much lower than the other two priors when magnitude is small, yet it overwhelms them as magnitude is getting larger. In addition, we also noticed the interaction effect disappears when intercepts have been contaminated. Data support the probability of the model with two main effects alone without any interaction effect (Table 2.14).

Unlike the BF, the uncertainty rate is much higher for Bayesian estimation. Figure 2.9 and Figure 2.10 display the rate across the levels of magnitude for method ROPE and ROPE_0 respectively. First, the uncertainty rate of both methods is much higher on loadings (M_{ROPE}= 0.857, SD_{ROPE} = 0.100; M_{ROPE_0} = 0.825, SD_{ROPE_0} = 0.089) than on intercepts (M_{ROPE}= 0.586, SD_{ROPE} = 0.417; M_{ROPE_0} = 0.475, SD_{ROPE_0} = 0.417). The uncertainty drops significantly on intercepts as magnitude of non-invariance increases. What is more, further Bayesian ANOVA (Table 2.15) shows the data is in favor of the single main effect *magnitude* model both on loadings ($BF_{\mathrm{M\_ROPE}}$ = 6.598, $BF_{\mathrm{M\_ROPE\_0}}$ = 9.282) and intercepts ($BF_{\mathrm{M\_ROPE}}$ = 9.410, $BF_{\mathrm{M\_ROPE\_0}}$ = 7.482). It indicates Bayesian estimation can locate the non-invariant items well especially when the magnitude is large. Finally, we noticed that data does not support the models with prior. It means uncertainty rate is quite similar among prior conditions.

Comparisons of Three Methods

To understand how well each method locates the non-invariance, we compare three methods by the criteria of both power and uncertainty rate. We will recommend the method with high power, rate of invariance and low uncertainty rate to common users in testing factorial invariance. Figure 2.11 and Figure 2.12 display the comparisons of power and uncertainty respectively. First, we find that BF functions better than Bayesian estimation on contaminated loadings, for its markedly higher power rate. However, on intercepts, the power of Bayesian estimation is superior to BF instead. ROPE_0 is higher (M_{ROPE_0} = 0.588, SD_{ROPE_0} = 0.400) than BF (M_{BF} = 0.442, SD_{BF} = 0.411) on average. In addition, we noticed that interaction effect is on loadings but not intercepts. The full model with both main and interaction effect is approved of data with the highest BF_M (Table 2.16).

Second, BF controls the uncertain rate much lower than Bayesian estimation, no matter where the non-invariance locates. On loadings, uncertain of BF (M_{BF} = 0.134, SD_{BF} = 0.1127) is far lower than ROPE (M_{ROPE} = 0.857, SD_{ROPE} = 0.100) or ROPE_0 ($M_{\mathrm{ROPE\_0}}$ = 0.825, $SD_{\mathrm{ROPE\_0}}$ = 0.089). On intercepts, though the differences of uncertainty between methods shrinks as the magnitudes increase, the gap is still large which BF holds the uncertainty significantly lower when magnitude is small. Moreover, the interaction effect has been supported by the data from both loadings and intercepts conditions (Table 2.17).

Third, we concentrate on the choice of prior on each method. Examining the conditions in which each prior applied for the methods to detect the non-invariance, BF is completely superior to the others. Figure 2.13 and Figure 2.14 presents the power and uncertainty rate. BF constantly holds the higher power and lower uncertainty rate. Though the power of ROPE_0 is slightly higher than BF under uniform conditions, its superiority stops at 0.4 magnitude of non-invariance. As the size of non-invariance continuously expands, BF re-gains the higher power and remains much higher. Moreover, we find that no matter what prior BF has chosen, its power rate (Figure 2.15) usually sustains the highest value (except with uniform prior on intercept). Furthermore, its uncertainty rate holds back to be the lowest among all methods (Figure 2.16). Both ROPE and ROPE_0 yield much higher uncertainty rate, even though they can quite successfully discover the non-invariance on intercept.

Finally, BF is exclusively better than Bayesian estimation in detecting the invariant items from the contaminated variables. Figure 2.17 shows the rate of which correctly identified invariance as well as the uncertainty rate of invariant parameters. On the left side, the rate of Bayesian estimation seems to be missing, but they are indeed zero across prior conditions. On the right side, the uncertainty rate of Bayesian estimation keeps very high. It means that Bayesian estimation fails to detect the invariant items.

A Pedagogical Example

To demonstrate the empirical application of Bayesian method, we use BF, ROPE and ROPE_0 to locate the non-invariance on the same data from Study I (N = 12,811) -- Psychological Wellbeing of Children of Rural-to-Urban Migrant Workers in China. The measurement chosen for this demonstration is from the Revised Child Anxiety and Depression Scale (RCADS, Chorpita, Yim, Moffitt, Umemoto & Francis, 2000). This self-report scale contains 47 items in total. However, only 18 items relate to generalized anxiety are used here for demonstration. Responses are scored on a Likert-scale of 1 to 4, corresponding to “Never”, “Sometimes”, “Quite Often”, and “Always”. The Cronbach’s α is 0.897 in this sample.

There are 7,356 male (57.4%) and 5,455 female (42.6%) child respondents in this sample. A two-group CFA is fitted to data, using the 7th variable as the reference indicator. We follow the same procedures in simulation to identify the model and put the Cauchy prior (0, 1) to use in Bayesian methods: Bayes factor, ROPE and ROPE_0. The parameter D_{ij} is computed as the parameter difference between groups. Three MCMC chains run 10,000 iterations after 500 burins to get the posterior distribution of parameter D_{ij}. MCMC is carried out for both loadings and intercepts simultaneously. Based on the posterior, both Bayes factor and 95% HDI are easily obtained. The Bayes factor is computed by Savage-Dickey density ratio and 3 as the cut-off value.

To achieve ROPE and ROPE_0, we also follow the procedures in our simulation study. For example, if 95% HDI completely excludes the interval between [-0.1, 0.1], we believe that non-invariance exists on the parameter. However, if 95% HDI partially overlaps with the interval but contains zero point, both method of “ROPE” and “ROPE_0” takes it inconclusively. While, if 95% HDI overlaps with the interval in which the point of zero excluded, the method “ROPE_0” believes the non-invariance is still on the parameters. Finally, if 95% HDI has been entirely contained within the interval, both methods agree to accept the null that invariance has been sustained.

Table 2.18 summarizes the results of three methods. We notice that they do not agree with each other in most of the time. According to BF, none of the factor loadings are invariant. BF_{01} produces the small values in which data is in favor of supporting the alternative hypothesis. However, this conclusion is not verified by either ROPE or ROPE_0. Except for item 10, 95% HDI contains the value zero as well as the range [-0.1, 0.1]. Therefore, both methods of Bayesian estimation are able to accept or reject null for the rest 16 loadings. On item 10, they agreed it to be non-invariant across groups. For the decisions on intercepts, Bayes factor accepts them to be invariance (except item 10), since BF_{01} values are larger than 3. Yet, the two methods of Bayesian estimation still fail to get any concrete conclusions. The intercept of item 10 is also non-invariant, as both Bayes factor and Bayesian estimation show the evidences for the alternative hypothesis.

In summary, regarding Bayes factor as a better method in simulation study, we decide to adopt its conclusions in current pedagogical example. None of the factor loadings are invariant, and most of intercepts (except item 10) are invariant between groups.

Discussion

The ability to locate the non-invariance ahead would notably benefit empirical users before they correctly conduct the partial invariance test. It can also aid in the search for the potential causality to non-invariance. Most methods applied in this area however, were monopolistically from the traditional Frequentist. The unavoidable defects of NHST prevent them from accepting the null in which parameters are invariant. Furthermore, they also suffer from the large sample size fallacy. The employment of new methods thereby becomes necessary. In the present study, we introduced the innovative approaches from Bayesian perspective. Bayesian estimation and Bayes factor were particularly applied to locate the sources of non-invariance. The Bayesian estimation is a general category for two subsume methods: ROPE and ROPE_0. Using the Gibbs sampling to run MCMC, Bayesian estimation can summarize the posterior distribution of cross-group parameter differences in terms of 95% HDI. Based on the relationship between 95% HDI and ROPE, both methods make decisions for hypothesis test. Depending on the same posterior distribution, we used the approach called Savage-Dickey density ratio to calculate Bayes factor. After performing the suggested cut-off value three, we decide to accept, reject or keep uncertain for hypothesis.

Our simulation study revealed that Bayes factor functions generally superior than Bayesian estimation. First, it yields higher power but lower uncertainty rate in most conditions. Particularly, as the magnitude of non-invariance increases, its power rate improves more rapidly and still maintains a low level of uncertainty. In addition, it is better to control the uncertainty rate under some circumstances in which non-invariance locates on loadings or the application of uniform prior. The power of BF essentially represents the features of *shrinkage* estimation[7]. That is the estimates are more likely towards value zero when a small observed effect size corresponds well with the null hypothesis. Yet, when the effect size becomes large (e.g., large magnitude), it will heavily impact the estimation by increasing the likelihood to accept the alternative.

Second, being consistent with previous literatures, BF is highly sensitive to the choices of priors. The values of BF vary with different priors. We chose three different non-informative priors: Normal, Cauchy and Uniform, because each of them represents some uniqueness of probability distribution. For example, Cauchy distribution has the longest tails on two sides. Uniform has the lowest density around zero, though it is limited within a certain probability range. Comparing to other two priors, the results of Cauchy provide higher power and lower uncertainty. Though, its density around null is higher than both Uniform and Normal priors, its heavy tails help it be less informative. In other words, BF is in favor of Cauchy to detect the non-invariance.

Third, BF is able to distinguish the invariant item much more accurately from those contaminated ones. It produces the extraordinarily high power rate with a well-controlled uncertainty rate. It will benefit researchers to accept the null when parameters are invariant. Differing from NHST that fail to reject null hypothesis has been mistakenly taken as an evidence for parameters' invariance, Bayes factor (BF_{01} > 3, or BF_{10} < 1/3) can indeed confirm the cross-group parameter invariance. Though most critiques about BF are its cut-off values, the current study still applied value three and obtained satisfactory results. However, these concerns about cut-off values are completely understandable from methodological perspectives. That is why we should not take BF as the only way of Bayesian application. Instead, both Bayes factor and Bayesian estimation should be applied like our pedagogical example. Ideally, the results from both are consistent. If not, we suggest to accept the BF's results alone.

On the perspective of Bayesian estimation, we used the methods ROPE and ROPE_0 to test the hypothesis of item parameters. The main distinction between the two methods is whether the value zero has been included in the range of 95% HDI, when 95% HDI overlaps with ROPE. If yes, both methods regard it as part of uncertain situation. If not, the method ROPE_0 takes the corresponding item to be non-invariant, while method ROPE still considers it to be uncertainty. For the current study, both ROPE and ROPE_0 represent the similar patterns on power and uncertainty rate. They share several characteristics including the insensitivity on the choice of prior. Consistent to the previous literatures, both power and uncertainty rate do not show a statistical difference among three non-informative priors. Second, even though the power is higher than Bayes factor under several conditions such as the non-invariance on intercepts, it is much lower in most cases. Due to the high posterior variance on item differences, the range of 95% HDI becomes much wider than ROPE, leading to the high uncertainty rate. Especially when the sample size is small, and the observed data cannot sufficiently provide the useful information, the posterior variations would remain high.

Future studies are still called for the area of factorial invariance test. For methodological researchers, a comparison between Bayesian and Frequentist approaches on testing the factorial invariance will be necessary. It should include either the *largest modification index* or *forward confidence interval* as the representative methods of frequentist. Therefore, a broader picture about the pros and cons for two sides will be clearly provided. For the empirical researchers, however, the most challenge is the applications of statistical packages in Bayesian area. It is now considerably difficult for common users to apply a customized model on JAGS. Further studies from both software developments and generalized practical utilization are needed.

Summary

Based on the findings in both Study I and Study II, a few suggestions may be offered to researchers. First, it is not wise to use *MaxL* to identify reference indicator. Although this approach could perform equally under certain conditions, it is impractical to identify those conditions in empirical data analysis. In addition, *MaxL* could behave poorly in large samples due to the sensitivity of LR test to sample size. Second, $Min\chi^2$ and BSEM are both recommended for empirical studies; however, different theoretical backgrounds are required for their implementation. While $Min\chi^2$ involves fitting a series of multiple-group CFA models and computing LR statistics for each individual item, BSEM is implemented through fitting a single model for identifying invariant and non-invariant items simultaneously (Shi, et al., 2017). In addition, we recommend methodological researchers to consider the direction of parameter differences as a studied variable involving simulation of multiple-group CFA models; otherwise the results could be cofounded or misleading. Furthermore, for the purpose of locating the non-invariance, it is recommended to take Bayes factor into account. With the non-informative of Cauchy prior, its superiority is high accurate to detect non-invariant item parameters. Furthermore, Bayes factor is able to distinguish the invariant items from these contaminated ones. Finally, the anticipation of user-friendly software packages would greatly improve further development in Bayesian methods.

Table 1.1: Power Rates of Selecting a Correct Reference Indicator in the Simulation Study

LO	PE	MA	SS	AR	Positive			Negative			Mix		
					MaxL	Min*χ2*	BSEM	MaxL	Min*χ2*	BSEM	MaxL	Min*χ2*	BSEM
Factor Loading	20%	0.2	100	0.80	0.19	0.95	0.95	1.00	0.96	0.95	0.65	0.99	0.98
			200	0.80	0.44	0.99	0.98	1.00	0.99	1.00	0.88	1.00	1.00
			500	0.80	0.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.4	100	0.80	0.85	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00
			200	0.80	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			500	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	40%	0.2	100	0.60	0.01	0.70	0.73	1.00	0.79	0.76	0.36	0.95	0.96
			200	0.60	0.01	0.78	0.79	1.00	0.90	0.90	0.71	1.00	0.99
			500	0.60	0.38	0.89	0.84	1.00	0.98	0.99	1.00	1.00	1.00
		0.4	100	0.60	0.06	0.77	0.78	1.00	0.99	0.99	0.96	1.00	1.00
			200	0.60	0.45	0.83	0.79	0.98	0.99	1.00	1.00	1.00	1.00
			500	0.60	0.07	0.90	0.80	0.25	0.99	1.00	1.00	1.00	1.00
Intercept	20%	0.3	100	0.80	0.82	1.00	0.99	0.97	0.99	0.99	0.98	1.00	1.00
			200	0.80	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			500	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
		0.6	100	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			200	0.80	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
			500	0.80	0.90	1.00	1.00	0.99	1.00	1.00	1.00	1.00	1.00
	40%	0.3	100	0.60	0.33	0.84	0.84	0.88	0.87	0.86	0.93	1.00	0.99
			200	0.60	0.49	0.92	0.92	0.95	0.94	0.92	1.00	1.00	1.00
			500	0.60	0.62	0.96	0.96	0.77	0.98	0.99	1.00	1.00	1.00
		0.6	100	0.60	0.72	0.95	0.94	0.93	0.98	0.98	1.00	1.00	1.00
			200	0.60	0.15	0.97	0.97	0.27	0.99	0.99	1.00	1.00	1.00
			500	0.60	0.00	0.94	0.99	0.00	0.95	0.99	1.00	1.00	1.00

Note: PE = Percentage of Non-invariance; LO = Location of Non-invariance; MA = Magnitude of Non-invariance; SS = Sample Size; AR = power rates of selecting a random item as RI.

Table 1.2: Effects of Studied Variables on Power Rates in the Simulation Study.

	ANOVA 1			ANOVA 2		
	df	*F*	*p*	*df*	*F*	*p*
Location (LO)	1	3.297	0.071	1	11.736	0.001
Percentage (PE)	1	33.608	<.001	1	119.617	<.001
Magnitude (MA)	1	0.690	0.407	1	2.455	0.120
Direction (DI)	2	19.623	<.001	2	69.842	<.001
SampleSize (SS)	2	0.583	0.559	2	2.074	0.131
Method (ME)	2	25.507	<.001	2	90.782	<.001
ME × MA				2	0.232	0.794
ME × LO				2	1.198	0.306
ME × PE				2	37.235	<.001
ME × DI				4	28.154	<.001
ME × SS				4	0.215	0.930
PE × MA				1	2.794	0.097
PE × LO				1	0.299	0.585
PE × DI				2	36.894	<.001
PE × SS				2	0.722	0.488
LO × MA				1	10.055	0.002
LO × DI				2	12.984	<.001
LO × SS				2	5.464	0.005
DI × MA				2	3.946	0.022
DI × SS				4	2.825	0.028
MA × SS				2	36.894	<.001
ME × MA × PE				2	9.400	<.001
ME × MA × LO				2	7.056	0.001
ME × MA × DI				4	7.964	<.001
ME × MA × SS				4	7.642	<.001
ME × DI × PE				4	9.840	<.001
ME × DI × LO				4	5.529	<.001
ME × DI × SS				8	3.779	0.001
ME × SS × PE				4	4.060	0.004
ME × SS × LO				4	3.000	0.022
ME × LO × PE				2	1.638	0.199
LO × PE × DI				2	3.506	0.033
LO × PE × MA				1	0.223	0.638
LO × PE × SS				2	0.721	0.489
LO × MA × DI				2	0.291	0.748
LO × MA × SS				2	1.604	0.206
LO × DI × SS				4	0.640	0.635
PE × MA × DI				2	2.151	0.121
PE × MA × SS				2	4.322	0.016
PE × DI × SS				4	0.973	0.426
MA × DI × SS				4	1.062	0.379
residuals	206			110		

Table 1.3: The Interaction Effect of Power between Methods and Directions

			Positive			Negative			Mix	
		df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
PE	20%	2	4.900	0.008	2	<.001	0.999	2	0.350	0.707
	40%	2	74.79	<.001	2	8.030	<.001	2	1.440	0.241
SS	N = 100	2	14.090	<.001	2	0.070	0.932	2	1.520	0.221
	N = 200	2	11.840	<.001	2	0.500	0.608	2	0.220	0.803
	N = 500	2	9.940	<.001	2	4.980	0.008	2	0.000	1.000
MA	small	2	21.530	<.001	2	0.030	0.966	2	1.880	0.155
	large	2	15.870	<.001	2	5.830	0.004	2	0.000	1.000
LO	loadings	2	27.300	<.001	2	0.120	0.883	2	1.840	0.162
	intercepts	2	12.390	<.001	2	3.650	0.028	2	0.010	0.993

Note: PE = Percentage of Non-invariance; SS = Sample Size; MA = Magnitude of Non-invariance; LO = Location of Non-invariance.

Table 1.4: Simple Effect of Power for Methods Comparisons on Directions

					Simple Effect											
		Methods			Positive				Negative				Mix			
		Comparison			*Diff*	*t*	*p*	*Adj p*	*Diff*	*t*	*p*	*Adj p*	*Diff*	*t*	*p*	*Adj p*
PE	20%	MaxL	-	Minχ^2	-0.153	-2.730	0.007	0.063	0.002	0.030	0.976	1.000	-0.041	-0.730	0.466	1.000
		MaxL	-	BSEM	-0.151	-2.700	0.008	0.069	0.002	0.030	0.976	1.000	-0.040	-0.720	0.475	1.000
		Minχ^2	-	BSEM	0.002	0.030	0.976	1.000	<.001	0.000	1.000	1.000	0.001	0.010	0.988	1.000
	40%	MaxL	-	Minχ^2	-0.597	-10.670	<.001	<.001	-0.193	-3.460	0.001	0.006	-0.083	-1.470	0.142	1.000
		MaxL	-	BSEM	-0.588	-10.520	<.001	<.001	-0.195	-3.490	0.001	0.005	-0.082	-1.460	0.146	1.000
		Minχ^2	-	BSEM	0.008	0.150	0.882	1.000	-0.002	-0.030	0.976	1.000	0.001	0.010	0.988	1.000
SS	N=100	MaxL	-	Minχ^2	-0.404	-4.580	<.001	<.001	0.025	0.280	0.777	1.000	-0.134	-1.520	0.131	1.000
		MaxL	-	BSEM	-0.406	-4.610	<.001	<.001	0.031	0.350	0.723	1.000	-0.133	-1.500	0.134	1.000
		Minχ^2	-	BSEM	-0.003	-0.030	0.997	1.000	0.006	0.070	0.944	1.000	0.001	0.010	0.989	1.000
	N=200	MaxL	-	Minχ^2	-0.374	-4.240	<.001	<.001	-0.076	-0.870	0.388	1.000	-0.051	-0.580	0.561	1.000
		MaxL	-	BSEM	-0.369	-4.190	<.001	<.001	-0.076	-0.870	0.388	1.000	-0.050	-0.570	0.571	1.000
		Minχ^2	-	BSEM	0.005	0.060	0.955	1.000	0.000	0.000	1.000	1.000	0.001	0.010	0.989	1.000
	N=500	MaxL	-	Minχ^2	-0.346	-3.930	<.001	0.001	-0.236	-2.680	0.008	0.072	<.001	0.000	1.000	1.000
		MaxL	-	BSEM	-0.334	-3.790	<.001	0.002	-0.245	-2.780	0.006	0.054	<-.001	0.000	1.000	1.000
		Minχ^2	-	BSEM	0.013	0.140	0.887	1.000	-0.009	-0.100	0.921	1.000	<-.001	0.000	1.000	1.000
MA	small	MaxL	-	Minχ^2	-0.402	-5.700	<.001	<.001	0.014	0.200	0.841	1.000	-0.119	-1.690	0.092	0.831
		MaxL	-	BSEM	-0.400	-5.670	<.001	<.001	0.018	0.250	0.804	1.000	-0.118	-1.670	0.097	0.873
		Minχ^2	-	BSEM	0.003	0.040	0.972	1.000	0.003	0.050	0.962	1.000	0.002	0.020	0.981	1.000
	large	MaxL	-	Minχ^2	-0.348	-4.930	<.001	<.001	-0.206	-2.920	0.004	0.035	-0.004	-0.060	0.953	1.000
		MaxL	-	BSEM	-0.340	-4.830	<.001	<.001	-0.211	-2.990	0.003	0.028	-0.004	-0.060	0.953	1.000
		Minχ^2	-	BSEM	0.008	0.110	0.915	1.000	-0.005	-0.070	0.944	1.000	<-.001	0.000	1.000	1.000
LO	loadings	MaxL	-	Minχ^2	-0.451	-6.490	<.001	<.001	-0.030	-0.430	0.666	1.000	-0.116	-1.670	0.097	0.874
		MaxL	-	BSEM	-0.438	-6.310	<.001	<.001	-0.030	-0.430	0.666	1.000	-0.115	-1.650	0.100	0.896
		Minχ^2	-	BSEM	0.013	0.180	0.857	1.000	<.001	0.000	1.000	1.000	0.001	0.010	0.990	1.000
	intercepts	MaxL	-	Minχ^2	-0.298	-4.290	<.001	<.001	-0.162	-2.330	0.021	0.189	-0.008	-0.110	0.914	1.000
		MaxL	-	BSEM	-0.301	-4.330	<.001	<.001	-0.163	-2.350	0.020	0.178	-0.007	-0.100	0.924	1.000
		Minχ^2	-	BSEM	-0.003	-0.040	0.971	1.000	-0.002	-0.020	0.981	1.000	0.001	0.010	0.990	1.000

Note: Adj *p* uses Bonferroni correction for familywise error rate. PE = Percentage of Non-invariance; SS = Sample Size; MA = Magnitude of Non-invariance; LO = Location of Non-invariance.

Table 1.5: The Interaction Effect of Power between Methods and Magnitudes

		Small Magnitude			Large Magnitude		
		df	*F*	*p*	*df*	*F*	*p*
PE	20%	2	2.330	0.100	2	0.050	0.956
	40%	2	9.400	<.001	2	23.67	<.001
SS	N = 100	2	5.980	0.003	2	0.990	0.374
	N = 200	2	3.020	0.051	2	2.630	0.074
	N = 500	2	0.820	0.441	2	9.060	<.001
DR	Positive	2	21.530	<.001	2	15.870	<.001
	Negative	2	0.030	0.966	2	5.830	0.004
	Mix	2	1.880	0.155	2	0.000	0.998
LO	loadings	2	8.600	<.001	2	3.750	0.025
	intercepts	2	1.480	0.230	2	7.050	0.001

Note: PE = Percentage of Non-invariance; SS = Sample Size; DR = Direction; LO = Location of Non-invariance.

Table 1.6: Simple Effect of Power for Methods Comparisons on Magnitude

					Simple Effect							
		Methods			Small Magnitude				Large Magnitude			
		Comparison			*Diff*	*t*	*p*	*Adj p*	*Diff*	*t*	*p*	*Adj p*
PE	20%	MaxL	-	Minχ^2	-0.112	-1.880	0.061	0.367	-0.016	-0.260	0.794	1.000
		MaxL	-	BSEM	-0.111	-1.850	0.065	0.391	-0.016	-0.260	0.794	1.000
		Minχ^2	-	BSEM	0.002	0.030	0.978	1.000	<.001	0.000	1.000	1.000
	40%	MaxL	-	Minχ^2	-0.223	-3.780	0<.001	0.001	-0.356	-5.970	<.001	<.001
		MaxL	-	BSEM	-0.222	-3.730	<.001	0.002	-0.354	-5.940	<.001	<.001
		Minχ^2	-	BSEM	0.003	0.060	0.956	1.000	0.002	0.030	0.978	1.000
SS	N=100	MaxL	-	Minχ^2	-0.243	-3.020	0.003	0.017	-0.098	-1.220	0.225	1.000
		MaxL	-	BSEM	-0.240	-2.970	0.003	0.020	-0.098	-1.220	0.225	1.000
		Minχ^2	-	BSEM	0.003	0.040	0.967	1.000	<.001	0.000	1.000	1.000
	N=200	MaxL	-	Minχ^2	-0.173	-2.140	0.034	0.203	-0.162	-2.000	0.047	0.279
		MaxL	-	BSEM	-0.171	-2.120	0.036	0.213	-0.159	-1.970	0.050	0.300
		Minχ^2	-	BSEM	0.002	0.020	0.984	1.000	0.003	0.030	0.975	1.000
	N=500	MaxL	-	Minχ^2	-0.091	-1.130	0.262	1.000	-0.298	-3.690	<.001	0.002
		MaxL	-	BSEM	-0.088	-1.090	0.275	1.000	-0.298	-3.690	<.001	0.002
		Minχ^2	-	BSEM	0.003	0.030	0.975	1.000	<.001	0.000	1.000	1.000
DI	Positive	MaxL	-	Minχ^2	-0.402	-5.700	<.001	<.001	-0.348	-4.930	<.001	<.001
		MaxL	-	BSEM	-0.399	-5.670	<.001	<.001	-0.340	-4.830	<.001	<.001
		Minχ^2	-	BSEM	0.003	0.040	0.972	1.000	0.008	0.110	0.915	1.000
	Negative	MaxL	-	Minχ^2	0.014	0.200	0.841	1.000	-0.206	-2.920	0.004	0.023
		MaxL	-	BSEM	0.018	0.250	0.804	1.000	-0.211	-2.990	0.003	0.019
		Minχ^2	-	BSEM	0.003	0.050	0.962	1.000	-0.005	-0.070	0.944	1.000
	Mix	MaxL	-	Minχ^2	-0.119	-1.690	0.092	0.554	-0.004	-0.060	0.953	1.000
		MaxL	-	BSEM	-0.118	-1.670	0.100	0.582	-0.004	-0.060	0.953	1.000
		Minχ^2	-	BSEM	0.002	0.020	0.981	1.000	<.001	0.000	1.000	1.000
LO	loadings	MaxL	-	Minχ^2	-0.238	-3.610	<.001	0.002	-0.159	-2.420	0.017	0.099
		MaxL	-	BSEM	-0.236	-3.570	<.001	0.003	-0.153	-2.320	0.021	0.127
		Minχ^2	-	BSEM	0.003	0.040	0.967	1.000	0.006	0.090	0.926	1.000
	intercepts	MaxL	-	Minχ^2	-0.099	-1.510	0.133	0.799	-0.212	-3.220	0.002	0.009
		MaxL	-	BSEM	-0.097	-1.470	0.142	0.852	-0.217	-3.280	0.001	0.007
		Minχ^2	-	BSEM	0.002	0.030	0.973	1.000	-0.004	-0.070	0.946	1.000

Note: Adj *p* uses Bonferroni correction for familywise error rate. PE = Percentage of Non-invariance; SS = Sample Size; DI = Direction; LO = Location of Non-invariance.

Table 1.7: Values of Selection Index in Choosing RI by $MaxL$ and Minχ^2 in the Empirical Analysis

	Factor loadings	G_1 loadings	G_2 loadings	χ^2	df	$\Delta\chi^2$	p
Baseline				6590.815	304		
Item 1	0.300	0.306	0.291	6576.866	302	13.949	<0.001
Item 2	0.480	0.471	0.492	6588.843	302	1.972	0.373
Item 3	0.571	0.556	0.591	6582.942	302	7.873	0.020
Item 4	0.620	0.611	0.633	6587.198	302	3.617	0.164
Item 5	0.643	0.630	0.661	6586.149	302	4.666	0.097
Item 6	0.538	0.557	0.510	6578.007	302	12.808	0.002
Item 7	**0.700**	0.703	0.694	6590.235	302	**0.580**	**0.748**
Item 8	0.656	0.651	0.663	6586.430	302	4.385	0.112
Item 9	0.665	0.653	0.682	6586.734	302	4.081	0.130
Item 10	0.690	0.682	0.703	6577.774	302	13.041	0.002
Item 11	0.540	0.555	0.518	6568.084	302	22.731	<0.001
Item 12	0.491	0.499	0.478	6575.632	302	15.183	<0.001
Item 13	0.536	0.543	0.528	6508.893	302	81.922	<0.001
Item 14	0.425	0.435	0.411	6582.484	302	8.331	0.016
Item 15	0.625	0.630	0.618	6589.869	302	0.946	0.623
Item 16	0.608	0.600	0.621	6584.566	302	6.249	0.044
Item 17	0.598	0.605	0.586	6589.292	302	1.523	0.467
Item 18	0.481	0.484	0.476	6590.227	302	0.588	0.745

Note: The column “Factor loading” provides the loadings from the Baseline model in which they are constraint to be equal cross groups; The columns “G_1 loadings” and “G_2 loadings” display the free estimates of loadings for the reference (G_1) and focal (G_2) group respectively.

Table 1.8: Values of Selection Index in Choosing RI Using BSEM in the Empirical analysis

	$D_{\widehat{factor\_loading}}$ (SD)	$D_{\widehat{intercept}}$ (SD)	Δ_j
Item 1	0.011 (0.014)	0.044 (0.014)	3.929
Item 2	0.017 (0.015)	0.007 (0.015)	1.600
Item 3	0.027 (0.016)	0.023 (0.015)	3.221
Item 4	0.019 (0.016)	0.016 (0.015)	2.254
Item 5	0.024 (0.015)	0.01 (0.015)	2.267
Item 6	0.036 (0.016)	0.027 (0.015)	4.050
Item 7	0.005 (0.016)	0.005 (0.015)	**0.646**
Item 8	0.01 (0.016)	0.024 (0.015)	2.225
Item 9	0.023 (0.016)	0.012 (0.015)	2.238
Item 10	0.017 (0.016)	0.037 (0.015)	3.529
Item 11	0.03 (0.013)	0.039 (0.013)	5.308
Item 12	0.018 (0.015)	0.041 (0.014)	4.129
Item 13	0.013 (0.015)	0.106 (0.015)	7.933
Item 14	0.019 (0.015)	0.03 (0.015)	3.267
Item 15	0.011 (0.014)	0.002 (0.014)	0.929
Item 16	0.016 (0.015)	0.022 (0.015)	2.533
Item 17	0.016 (0.015)	0.001 (0.015)	1.133
Item 18	0.007 (0.016)	0.007 (0.015)	0.904

Note: SD is standard deviation.

Table 2.1: The Rate of Bayes Factor to Reject Null Hypothesis

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.086	0.086	0.084	0.116	0.222	0.346	0.536	0.670	0.760
		200	0.080	0.082	0.086	0.150	0.352	0.568	0.738	0.814	0.830
		500	0.080	0.080	0.088	0.300	0.692	0.852	0.900	0.948	0.952
		unbalanced	0.074	0.076	0.082	0.206	0.560	0.846	0.974	0.994	0.998
Loading	Cauchy	100	0.130	0.144	0.160	0.288	0.506	0.740	0.870	0.958	0.980
		200	0.104	0.110	0.156	0.388	0.718	0.906	0.972	0.994	1.000
		500	0.098	0.142	0.220	0.720	0.948	0.964	0.986	0.986	0.990
		unbalanced	0.096	0.122	0.214	0.564	0.932	0.988	0.998	1.000	1.000
	Normal	100	0.118	0.122	0.116	0.192	0.398	0.632	0.824	0.940	0.980
		200	0.088	0.088	0.098	0.298	0.660	0.900	0.984	0.998	1.000
		500	0.110	0.128	0.188	0.692	0.974	0.994	0.998	1.000	1.000
		unbalanced	0.098	0.112	0.162	0.494	0.882	0.982	0.998	1.000	1.000
	uniform	100	0.006	0.006	0.006	0.010	0.016	0.038	0.130	0.27	0.482
		200	0.036	0.040	0.040	0.058	0.156	0.390	0.678	0.888	0.968
		500	0.020	0.020	0.026	0.198	0.662	0.962	1.000	1.000	1.000
		unbalanced	0.000	0.000	0.002	0.016	0.154	0.484	0.892	0.992	1.000
Intercept	Cauchy	100	0.002	0.000	0.004	0.014	0.044	0.126	0.268	0.448	0.664
		200	0.006	0.008	0.008	0.062	0.194	0.446	0.682	0.834	0.938
		500	0.012	0.024	0.078	0.546	0.936	0.986	0.998	1.000	1.000
		unbalanced	0.000	0.004	0.016	0.210	0.702	0.962	1.000	1.000	1.000
	Normal	100	0.002	0.004	0.002	0.024	0.040	0.130	0.274	0.450	0.664
		200	0.006	0.006	0.010	0.050	0.158	0.358	0.610	0.794	0.926
		500	0.010	0.008	0.050	0.456	0.898	0.972	0.990	0.998	1.000
		unbalanced	0.000	0.006	0.020	0.184	0.646	0.954	1.000	1.000	1.000

Table 2.2: The Rate of Bayes factor to Accept Null Hypothesis

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.908	0.888	0.866	0.730	0.542	0.324	0.190	0.118	0.064
		200	0.908	0.908	0.874	0.700	0.386	0.204	0.138	0.106	0.086
		500	0.918	0.906	0.866	0.426	0.122	0.068	0.042	0.024	0.010
		unbalanced	0.924	0.908	0.858	0.566	0.146	0.028	0.004	0.000	0.000
Loading	Cauchy	100	0.662	0.584	0.488	0.266	0.114	0.028	0.008	0.000	0.000
		200	0.754	0.684	0.510	0.214	0.036	0.008	0.000	0.000	0.000
		500	0.770	0.648	0.382	0.030	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.796	0.712	0.474	0.102	0.008	0.000	0.000	0.000	0.000
	Normal	100	0.716	0.706	0.674	0.436	0.182	0.070	0.018	0.002	0.000
		200	0.810	0.784	0.638	0.298	0.060	0.010	0.000	0.000	0.000
		500	0.822	0.736	0.496	0.058	0.000	0.000	0.002	0.000	0.000
		unbalanced	0.796	0.772	0.608	0.182	0.010	0.002	0.000	0.000	0.000
	uniform	100	0.992	0.990	0.988	0.974	0.960	0.810	0.628	0.448	0.220
		200	0.958	0.956	0.946	0.884	0.678	0.320	0.100	0.046	0.014
		500	0.980	0.976	0.948	0.618	0.114	0.004	0.000	0.000	0.000
		unbalanced	1.000	0.998	0.996	0.912	0.588	0.158	0.018	0.000	0.000
Intercept	Cauchy	100	0.856	0.912	0.876	0.744	0.484	0.200	0.052	0.010	0.000
		200	0.926	0.932	0.836	0.466	0.162	0.010	0.000	0.000	0.000
		500	0.890	0.870	0.568	0.072	0.004	0.000	0.000	0.000	0.000
		unbalanced	0.970	0.966	0.822	0.298	0.024	0.000	0.000	0.000	0.000
	Normal	100	0.904	0.944	0.920	0.782	0.510	0.236	0.054	0.004	0.000
		200	0.954	0.950	0.884	0.512	0.176	0.016	0.000	0.000	0.000
		500	0.938	0.900	0.622	0.078	0.002	0.000	0.000	0.000	0.000
		unbalanced	0.972	0.964	0.832	0.332	0.034	0.000	0.000	0.000	0.000

Table 2.3: Bayesian ANOVA on Power Rate of Bayes Factor between Prior and Magnitude

Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
PR	0.200	0.200	<.001	<.001	<.001	<.001	0.567	0.125	0.009	0.007
MA	0.200	0.200	<.001	0.828	<.001	**>10.000**	>10.000	>10.000	<.001	<.001
PR + MA	0.200	0.200	0.899	0.165	**>10.000**	0.791	>10.000	>10.000	0.986	0.874
PR +MA + PR×MA	0.200	0.200	0.101	0.006	0.448	0.026	>10.000	>10.000	0.967	1.303

Note: PR = Prior, MA = Magnitude

Table 2.4: The Interaction Effect of Bayes Factor between Magnitude and Sample Size

Models	P(M)			P(M|data)			BF_M			BF_{10}			Error%		
	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor
Null	0.200	0.200	0.200	<.001	<.001	<.001	<.001	<.001	<.001	1.000	1.000	1.000			
SS	0.200	0.200	0.200	<.001	<.001	<.001	<.001	<.001	<.001	1.627	0.933	1.114	<.001	<.001	<.001
MA	0.200	0.200	0.200	<.001	<.001	<.001	<.001	0.003	<.001	>10.000	>10.000	>10.000	<.001	<.001	<.001
SS + MA	0.200	0.200	0.200	0.306	0.833	0.780	1.765	**>10.000**	**>10.000**	>10.000	>10.000	>10.000	0.627	1.015	1.176
SS + MA + SS×MA	0.200	0.200	0.200	0.694	0.166	0.220	**9.065**	0.797	1.130	>10.000	>10.000	>10.000	0.832	0.566	0.657

Note: Uni = Uniform prior; Cau = Cauchy prior; Nor = Normal prior; SS = Sample Size, MA = Magnitude

Table 2.5: The Rate of ROPE to Reject Null Hypothesis

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.020	0.026	0.038	0.076	0.142	0.196	0.222	0.238	0.242
		200	0.030	0.032	0.042	0.128	0.204	0.262	0.282	0.282	0.282
		500	0.020	0.020	0.030	0.106	0.214	0.244	0.244	0.244	0.244
		unbalanced	0.022	0.03	0.044	0.124	0.208	0.230	0.232	0.234	0.234
Loading	Cauchy	100	0.024	0.024	0.026	0.038	0.090	0.140	0.194	0.208	0.224
		200	0.028	0.028	0.034	0.064	0.146	0.228	0.250	0.260	0.260
		500	0.022	0.022	0.024	0.112	0.222	0.246	0.246	0.246	0.246
		unbalanced	0.014	0.014	0.028	0.100	0.220	0.240	0.248	0.248	0.248
	Normal	100	0.022	0.020	0.020	0.044	0.082	0.126	0.188	0.208	0.224
		200	0.024	0.024	0.026	0.050	0.116	0.182	0.222	0.236	0.236
		500	0.028	0.028	0.030	0.102	0.232	0.268	0.268	0.268	0.268
		unbalanced	0.020	0.022	0.028	0.084	0.212	0.260	0.266	0.270	0.270
	uniform	100	0.012	0.016	0.020	0.048	0.122	0.310	0.546	0.752	0.902
		200	0.046	0.048	0.054	0.118	0.354	0.710	0.920	0.978	0.998
		500	0.020	0.020	0.034	0.276	0.818	0.992	1.000	1.000	1.000
		unbalanced	0.000	0.002	0.008	0.164	0.628	0.950	1.000	1.000	1.000
Intercept	Cauchy	100	0.000	0.000	0.002	0.012	0.044	0.118	0.240	0.402	0.628
		200	0.000	0.002	0.006	0.034	0.130	0.338	0.566	0.768	0.910
		500	0.010	0.012	0.018	0.200	0.742	0.968	0.994	1.000	1.000
		unbalanced	0.000	0.000	0.006	0.122	0.590	0.928	1.000	1.000	1.000
	Normal	100	0.000	0.002	0.000	0.022	0.040	0.130	0.274	0.462	0.682
		200	0.002	0.004	0.004	0.020	0.144	0.336	0.610	0.810	0.932
		500	0.006	0.006	0.014	0.214	0.752	0.968	0.996	1.000	1.000
		unbalanced	0.000	0.000	0.008	0.102	0.544	0.932	0.994	0.998	0.998

Table 2.6: The Rate of ROPE to Accept Null Hypothesis

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Loading	uniform	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Cauchy	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Normal	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Intercept	uniform	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Cauchy	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Normal	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2.7: The Rate of ROPE with Zero to Reject Null Hypothesis

							Magnitude				
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.032	0.044	0.084	0.140	0.198	0.220	0.244	0.246	0.250
		200	0.044	0.072	0.108	0.210	0.272	0.282	0.282	0.282	0.282
		500	0.034	0.060	0.110	0.222	0.244	0.244	0.244	0.244	0.244
		unbalanced	0.044	0.086	0.132	0.216	0.234	0.234	0.234	0.234	0.234
Loading	Cauchy	100	0.036	0.036	0.042	0.096	0.136	0.196	0.220	0.226	0.228
		200	0.036	0.048	0.078	0.158	0.220	0.258	0.260	0.260	0.260
		500	0.032	0.058	0.116	0.222	0.246	0.246	0.246	0.246	0.246
		unbalanced	0.030	0.062	0.108	0.226	0.244	0.248	0.248	0.248	0.248
	Normal	100	0.030	0.036	0.048	0.084	0.142	0.188	0.212	0.222	0.228
		200	0.038	0.034	0.056	0.130	0.210	0.230	0.236	0.236	0.236
		500	0.032	0.064	0.134	0.250	0.268	0.268	0.268	0.268	0.268
		unbalanced	0.042	0.052	0.106	0.214	0.266	0.270	0.270	0.270	0.270
	uniform	100	0.028	0.030	0.058	0.128	0.320	0.538	0.770	0.890	0.956
		200	0.058	0.072	0.126	0.402	0.694	0.938	0.988	1.000	0.998
		500	0.034	0.078	0.260	0.830	0.990	1.000	1.000	1.000	1.000
		unbalanced	0.018	0.050	0.154	0.660	0.964	0.998	1.000	1.000	1.000
Intercept	Cauchy	100	0.018	0.006	0.008	0.050	0.122	0.242	0.420	0.598	0.798
		200	0.010	0.012	0.038	0.166	0.340	0.558	0.770	0.896	0.974
		500	0.022	0.042	0.184	0.742	0.960	0.994	1.000	1.000	1.000
		unbalanced	0.014	0.024	0.122	0.586	0.934	0.996	1.000	1.000	1.000
	Normal	100	0.020	0.010	0.014	0.058	0.134	0.268	0.462	0.672	0.838
		200	0.006	0.014	0.022	0.158	0.336	0.594	0.792	0.944	0.982
		500	0.018	0.050	0.214	0.768	0.978	0.994	1.000	1.000	1.000
		unbalanced	0.014	0.026	0.128	0.584	0.952	0.998	0.998	0.998	0.998

Table 2.8: The Rate of ROPE with Zero to Accept Null Hypothesis

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Loading	Cauchy	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Normal	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	uniform	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Intercept	Cauchy	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Normal	100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		200	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		500	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
		unbalanced	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 2.9: The Interaction Effect of Bayesian Estimation Power between Magnitude and Prior

Method				ROPE						
Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
PR	0.200	0.200	<.001	<.001	<.001	<.001	0.116	0.142	0.007	0.007
MA	0.200	0.200	0.602	0.701	**6.061**	**9.360**	>10.000	>10.000	0.002	<.001
PR + MA	0.200	0.200	0.380	0.289	2.453	1.625	>10.000	>10.000	0.812	0.903
PR +MA + PR×MA	0.200	0.200	0.017	0.011	0.071	0.042	>10.000	>10.000	1.387	0.658
Method				ROPE with zero						
Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
PR	0.200	0.200	<.001	<.001	<.001	<.001	0.120	0.157	0.007	0.007
MA	0.200	0.200	0.705	0.655	**9.550**	**7.597**	>10.000	>10.000	<.001	<.001
PR + MA	0.200	0.200	0.284	0.333	1.586	1.995	>10.000	>10.000	0.818	0.958
PR +MA + PR×MA	0.200	0.200	0.011	0.012	0.046	0.049	>10.000	>10.000	0.929	2.173

Note: PR = Prior, MA = Magnitude.

Table 2.10: The Interaction Effect of Bayesian Estimation Power between Magnitude and Sample Size

Method	ROPE														
Models	P(M)			P(M|data)			BF_M			BF_{10}			Error%		
	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor
Null	0.200	0.200	0.200	<.001	<.001	<.001	0.002	0.002	0.001	1.000	1.000	1.000			
SS	0.200	0.200	0.200	<.001	<.001	<.001	<.001	0.001	<.001	0.162	0.630	0.582	<.001	<.001	<.001
MA	0.200	0.200	0.200	0.795	0.302	0.310	**15.514**	1.730	1.801	>10.000	>10.000	>10.000	0.004	0.010	<.001
SS + MA	0.200	0.200	0.200	0.194	0.650	0.642	0.960	**7.425**	**7.173**	>10.000	>10.000	>10.000	0.478	0.441	0.465
SS + MA + SS×MA	0.200	0.200	0.200	0.011	0.047	0.047	0.044	0.199	0.198	>10.000	>10.000	>10.000	0.513	0.428	0.474
Method	ROPE with zero														
Models	P(M)			P(M|data)			BF_M			BF_{10}			Error%		
	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor	Uni	Cau	Nor
Null	0.200	0.200	0.200	0.021	0.010	0.008	0.087	0.039	0.033	1.000	1.000	1.000			
SS	0.200	0.200	0.200	0.004	0.006	0.005	0.014	0.026	0.022	0.165	0.667	0.668	<.001	<.001	<.001
MA	0.200	0.200	0.200	0.800	0.381	0.377	**15.968**	2.461	2.419	>10.000	>10.000	>10.000	0.001	0.001	<.001
SS + MA	0.200	0.200	0.200	0.167	0.569	0.574	0.799	**5.271**	**5.385**	7.794	>10.000	>10.000	0.484	0.497	0.369
SS + MA + SS×MA	0.200	0.200	0.200	0.009	0.035	0.036	0.036	0.143	0.149	0.415	3.604	4.457	0.548	0.577	0.607

Note: Uni = Uniform prior; Cau = Cauchy prior; Nor = Normal prior; SS = Sample Size, MA = Magnitude

Table 2.11: The Uncertainty Rate of Bayes Factor

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.006	0.026	0.050	0.154	0.236	0.330	0.274	0.212	0.176
		200	0.012	0.010	0.040	0.150	0.262	0.228	0.124	0.080	0.084
		500	0.002	0.014	0.046	0.274	0.186	0.080	0.058	0.028	0.038
		unbalanced	0.002	0.016	0.060	0.228	0.294	0.126	0.022	0.006	0.002
Loading	Cauchy	100	0.208	0.272	0.352	0.446	0.380	0.232	0.122	0.042	0.020
		200	0.142	0.206	0.334	0.398	0.246	0.086	0.028	0.006	0.000
		500	0.132	0.210	0.398	0.250	0.052	0.036	0.014	0.014	0.010
		unbalanced	0.108	0.166	0.312	0.334	0.060	0.012	0.002	0.000	0.000
	Normal	100	0.166	0.172	0.210	0.372	0.420	0.298	0.158	0.058	0.020
		200	0.102	0.128	0.264	0.404	0.280	0.090	0.016	0.002	0.000
		500	0.068	0.136	0.316	0.250	0.026	0.006	0.000	0.000	0.000
		unbalanced	0.106	0.116	0.230	0.324	0.108	0.016	0.002	0.000	0.000
	uniform	100	0.002	0.004	0.006	0.016	0.024	0.152	0.242	0.282	0.298
		200	0.006	0.004	0.014	0.058	0.166	0.290	0.222	0.066	0.018
		500	0.000	0.004	0.026	0.184	0.224	0.034	0.000	0.000	0.000
		unbalanced	0.000	0.002	0.002	0.072	0.258	0.358	0.090	0.008	0.000
Intercept	Cauchy	100	0.142	0.088	0.120	0.242	0.472	0.674	0.680	0.542	0.336
		200	0.068	0.060	0.156	0.472	0.644	0.544	0.318	0.166	0.062
		500	0.098	0.106	0.354	0.382	0.060	0.014	0.002	0.000	0.000
		unbalanced	0.030	0.030	0.162	0.492	0.274	0.038	0.000	0.000	0.000
	Normal	100	0.094	0.052	0.078	0.194	0.450	0.634	0.672	0.546	0.336
		200	0.040	0.044	0.106	0.438	0.666	0.626	0.390	0.206	0.074
		500	0.052	0.092	0.328	0.466	0.100	0.028	0.010	0.002	0.000
		unbalanced	0.028	0.030	0.148	0.484	0.320	0.046	0.000	0.000	0.000

Table 2.12: The Uncertain Rate of ROPE

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.980	0.974	0.962	0.924	0.858	0.804	0.778	0.762	0.758
		200	0.970	0.968	0.958	0.872	0.796	0.738	0.718	0.718	0.718
		500	0.980	0.980	0.970	0.894	0.786	0.756	0.756	0.756	0.756
		unbalanced	0.978	0.970	0.956	0.876	0.792	0.770	0.768	0.766	0.766
Loading	Cauchy	100	0.976	0.976	0.974	0.962	0.910	0.860	0.806	0.792	0.776
		200	0.972	0.972	0.966	0.936	0.854	0.772	0.750	0.740	0.740
		500	0.976	0.978	0.976	0.888	0.778	0.754	0.754	0.754	0.754
		unbalanced	0.986	0.986	0.972	0.900	0.780	0.760	0.752	0.752	0.752
	Normal	100	0.978	0.980	0.980	0.956	0.918	0.874	0.812	0.792	0.776
		200	0.976	0.976	0.974	0.950	0.884	0.818	0.778	0.764	0.764
		500	0.970	0.970	0.970	0.898	0.768	0.732	0.732	0.732	0.732
		unbalanced	0.980	0.978	0.972	0.916	0.788	0.740	0.734	0.730	0.730
	uniform	100	0.988	0.984	0.980	0.952	0.878	0.690	0.454	0.248	0.098
		200	0.954	0.952	0.946	0.882	0.646	0.290	0.080	0.022	0.002
		500	0.980	0.980	0.966	0.724	0.182	0.008	0.000	0.000	0.000
		unbalanced	1.000	0.998	0.992	0.836	0.372	0.050	0.000	0.000	0.000
Intercept	Cauchy	100	1.000	1.000	0.998	0.988	0.956	0.882	0.760	0.598	0.372
		200	1.000	0.998	0.994	0.966	0.870	0.662	0.434	0.232	0.090
		500	0.990	0.988	0.982	0.800	0.258	0.032	0.006	0.000	0.000
		unbalanced	1.000	1.000	0.994	0.878	0.410	0.072	0.000	0.000	0.000
	Normal	100	1.000	0.998	1.000	0.978	0.960	0.870	0.726	0.538	0.318
		200	0.998	0.996	0.996	0.980	0.856	0.664	0.390	0.190	0.068
		500	0.994	0.994	0.986	0.786	0.248	0.032	0.004	0.000	0.000
		unbalanced	1.000	1.000	0.992	0.898	0.456	0.068	0.006	0.002	0.002

Table 2.13: The Uncertain Rate of ROPE with Zero

			Magnitude								
LO	Prior	SS	0	0.05	0.1	0.2	0.3	0.4	0.5	0.6	0.7
	uniform	100	0.968	0.956	0.916	0.860	0.802	0.780	0.756	0.754	0.750
		200	0.956	0.928	0.892	0.790	0.728	0.718	0.718	0.718	0.718
		500	0.966	0.940	0.890	0.778	0.756	0.756	0.756	0.756	0.756
		unbalanced	0.956	0.914	0.868	0.784	0.766	0.766	0.766	0.766	0.766
Loading	Cauchy	100	0.964	0.964	0.958	0.904	0.864	0.804	0.780	0.774	0.772
		200	0.964	0.952	0.922	0.842	0.780	0.742	0.740	0.740	0.740
		500	0.966	0.942	0.884	0.778	0.754	0.754	0.754	0.754	0.754
		unbalanced	0.970	0.938	0.892	0.774	0.756	0.752	0.752	0.752	0.752
	Normal	100	0.970	0.964	0.952	0.916	0.858	0.812	0.788	0.778	0.772
		200	0.962	0.966	0.944	0.870	0.790	0.770	0.764	0.764	0.764
		500	0.966	0.934	0.866	0.750	0.732	0.732	0.732	0.732	0.732
		unbalanced	0.958	0.948	0.894	0.786	0.734	0.730	0.730	0.730	0.730
	uniform	100	0.972	0.970	0.942	0.872	0.680	0.462	0.230	0.110	0.044
		200	0.942	0.928	0.874	0.598	0.306	0.062	0.012	0.000	0.002
		500	0.966	0.922	0.740	0.170	0.010	0.000	0.000	0.000	0.000
		unbalanced	0.982	0.950	0.846	0.340	0.036	0.002	0.000	0.000	0.000
Intercept	Cauchy	100	0.982	0.994	0.992	0.950	0.878	0.758	0.580	0.402	0.202
		200	0.990	0.988	0.962	0.834	0.660	0.442	0.230	0.104	0.026
		500	0.978	0.958	0.816	0.258	0.040	0.006	0.000	0.000	0.000
		unbalanced	0.986	0.976	0.878	0.414	0.066	0.004	0.000	0.000	0.000
	Normal	100	0.980	0.990	0.986	0.942	0.866	0.732	0.538	0.328	0.162
		200	0.994	0.986	0.978	0.842	0.664	0.406	0.208	0.056	0.018
		500	0.982	0.950	0.786	0.232	0.022	0.006	0.000	0.000	0.000
		unbalanced	0.986	0.974	0.872	0.416	0.048	0.002	0.002	0.002	0.002

Table 2.14: Bayesian ANOVA on Uncertainty Rate of Bayes Factor between Prior and Magnitude

Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	0.007	<.001	0.029	1.000	1.000		
PR	0.200	0.200	<.001	0.038	<.001	0.159	0.149	5.319	0.007	0.005
MA	0.200	0.200	0.007	0.072	0.028	0.312	>10.000	10.071	0.008	0.006
PR + MA	0.200	0.200	0.001	0.833	0.006	>**10.000**	>10.000	>10.000	0.640	2.197
PR +MA + PR×MA	0.200	0.200	0.992	0.049	>**10.000**	0.206	>10.000	6.819	0.759	1.034

Note: PR = Prior, MA = Magnitude

Table 2.15: Bayesian ANOVA on Uncertainty Rate of Bayesian Estimation between Prior and Magnitude

Method				ROPE						
Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
PR	0.200	0.200	<.001	<.001	<.001	<.001	0.112	0.126	0.007	0.007
MA	0.200	0.200	0.623	0.702	**6.598**	**9.410**	>10.000	>10.000	0.005	<.001
PR + MA	0.200	0.200	0.363	0.290	2.280	1.631	>10.000	>10.000	1.265	0.752
PR +MA + PR×MA	0.200	0.200	0.014	0.009	0.058	0.035	>10.000	>10.000	1.375	1.134

Method				ROPE with zero						
Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
PR	0.200	0.200	<.001	<.001	<.001	<.001	0.103	0.134	0.007	0.007
MA	0.200	0.200	0.699	0.652	**9.282**	**7.482**	>10.000	>10.000	<.001	0.011
PR + MA	0.200	0.200	0.292	0.339	1.652	2.048	>10.000	>10.000	1.031	0.838
PR +MA + PR×MA	0.200	0.200	0.009	0.010	0.036	0.039	>10.000	>10.000	0.626	0.849

Note: PR = Prior, MA = Magnitude

Table 2.16: Compare the Power Rate among Methods

Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
ME	0.200	0.200	<.001	<.001	<.001	<.001	>10.000	0.925	<.001	0.006
MA	0.200	0.200	<.001	0.001	<.001	0.005	>10.000	>10.000	<.001	<.001
ME + MA	0.200	0.200	<.001	0.973	<.001	**>10.000**	>10.000	>10.000	1.809	0.744
ME +MA + ME×MA	0.200	0.200	1.000	0.026	**>10.000**	0.108	>10.000	>10.000	0.909	1.006

Note: ME = Method, MA = Magnitude

Table 2.17: Compare the Uncertainty Rate among Methods

Models	P(M)		P(M|data)		BF_M		BF_{10}		Error%	
	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts	loadings	intercepts
Null model	0.200	0.200	<.001	<.001	<.001	<.001	1.000	1.000		
ME	0.200	0.200	<.001	<.001	<.001	<.001	>10.000	>10.000	<.001	0.012
MA	0.200	0.200	<.001	<.001	<.001	<.001	0.242	>10.000	<.001	<.001
ME + MA	0.200	0.200	<.001	<.001	<.001	<.001	>10.000	>10.000	0.767	1.094
ME +MA + ME×MA	0.200	0.200	1.000	1.000	**>10.000**	**>10.000**	>10.000	>10.000	0.883	1.192

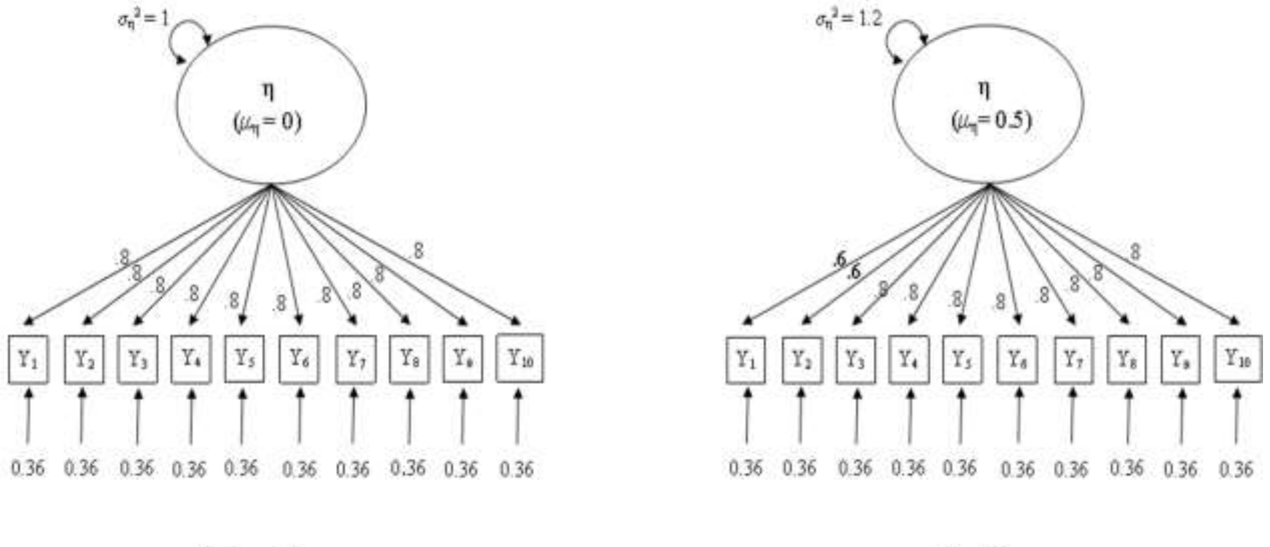
Note: ME = Method, MA = Magnitude

Table 2.18: Apply Bayes Factor and Bayesian Estimation in the Empirical Analysis

	Item	1	2	3	4	5	6	7	8	9
	BF_{01}	0.002	<0.001	<0.001	<0.001	<0.001	<0.001	RI	<0.001	<0.001
Loadings	lower	-5.696	-9.425	-10.699	-11.771	-12.153	-9.630	RI	-12.241	-12.481
	upper	1.705	2.804	3.229	3.542	3.641	2.960	RI	3.696	3.787
	BF_{01}	3.603	13.205	5.341	6.201	10.260	5.015	RI	4.913	6.471
Intercepts	lower	-0.232	-0.268	-0.347	-0.361	-0.338	-0.317	RI	-0.386	-0.390
	upper	0.576	1.055	1.166	1.295	1.378	1.043	RI	1.339	1.373
	Item	10	11	12	13	14	15	16	17	18
	BF_{01}	N/A	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Loadings	lower	1.927	-9.726	-9.375	-10.282	-8.230	-11.701	-12.003	-11.489	-9.424
	upper	3.393	2.986	2.788	3.047	2.433	3.556	3.528	3.399	2.789
	BF_{01}	N/A	3.661	18.192	4.694	21.476	9.244	13.605	9.324	16.047
Intercepts	lower	-0.373	-0.336	-0.222	-0.162	-0.204	-0.334	-0.325	-0.320	-0.274
	upper	0.039	1.043	1.094	1.282	0.951	1.317	1.370	1.303	1.047

Note: the loading and intercept of Item 10 was not convergent to calculate BF_{01}, because the value was extremely small.

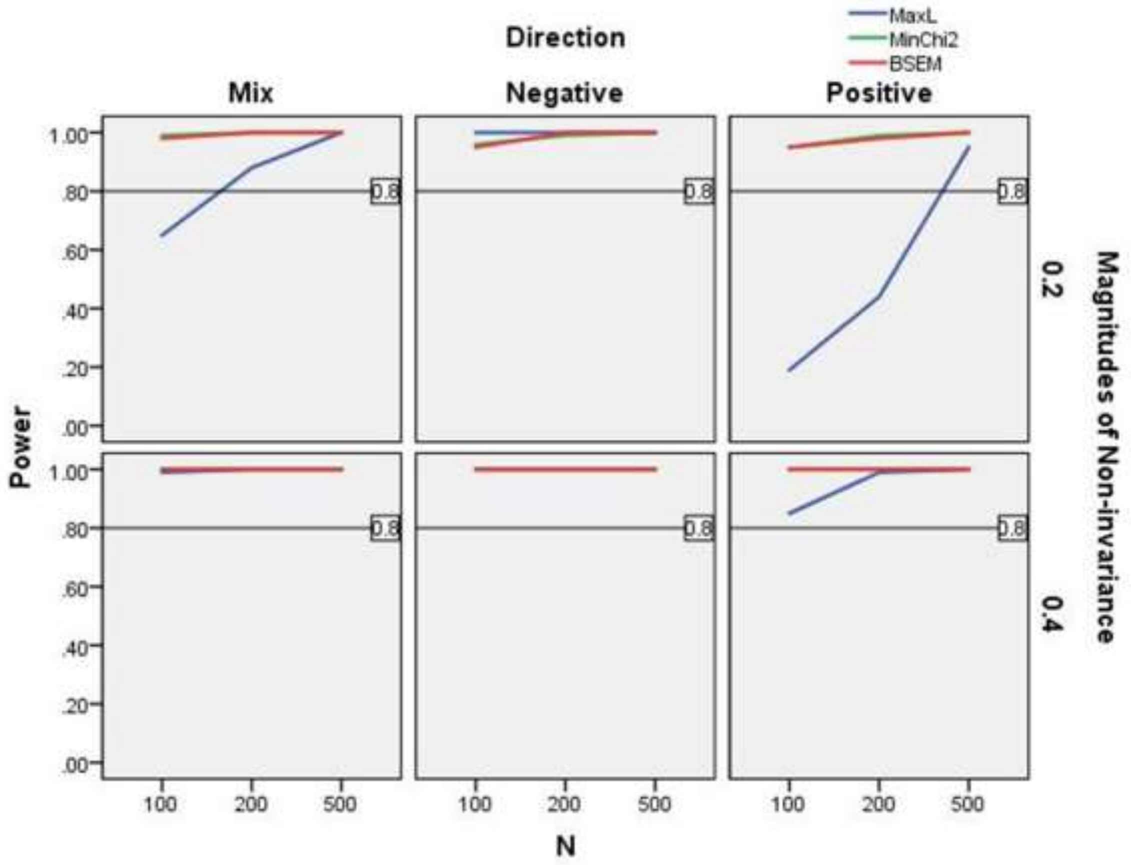
Figure 1.1: An Example of Two-group CFA Model for Data Simulation



Reference Group

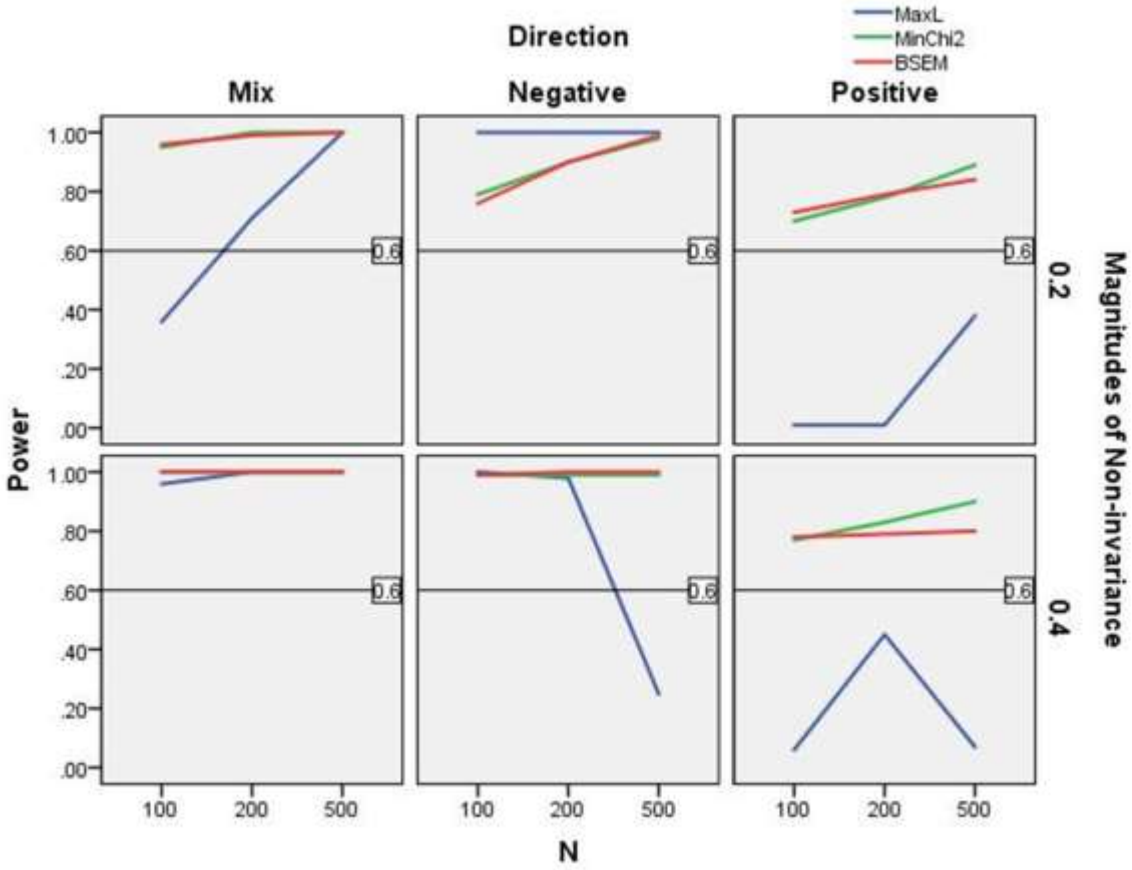
Focal Group

Figure 1.2: 20% Factor Loadings Have Non-invariance



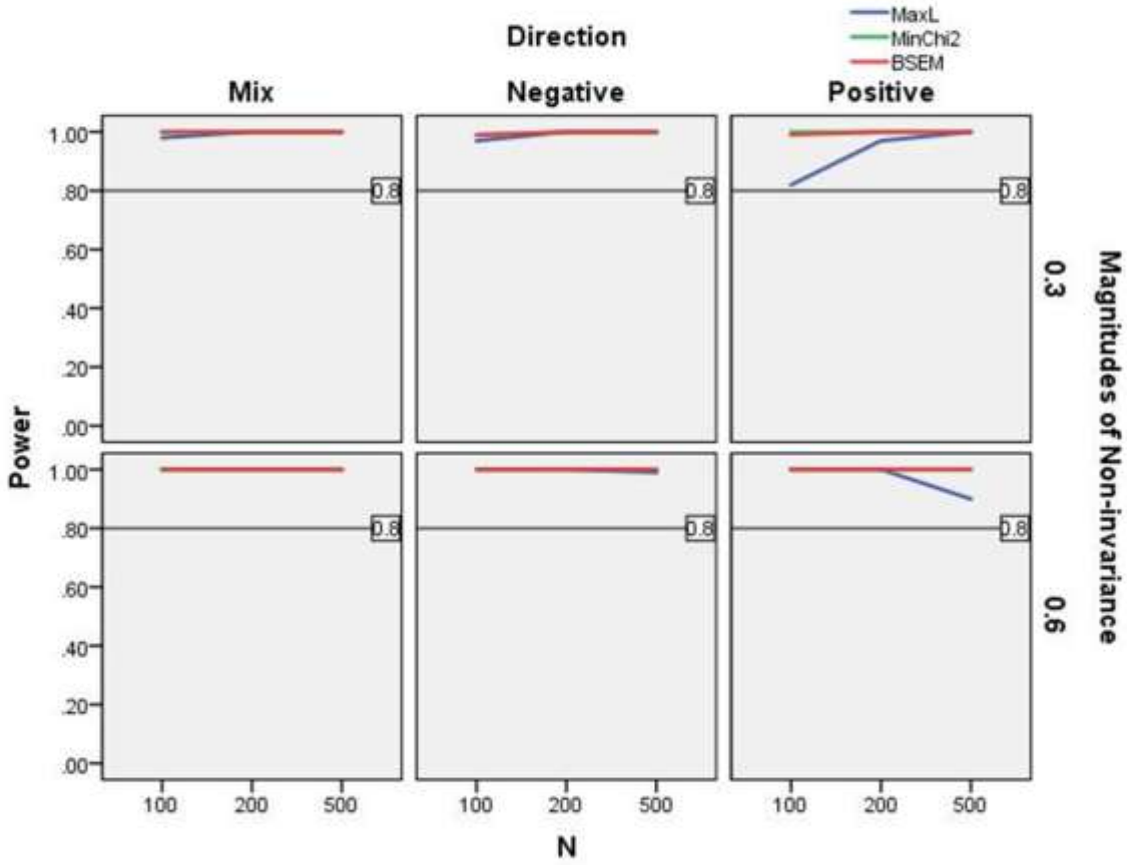
Note: The reference line indicates the power rates as randomly selecting an item as RI.

Figure 1.3: 40% Factor Loadings Have Non-Invariance



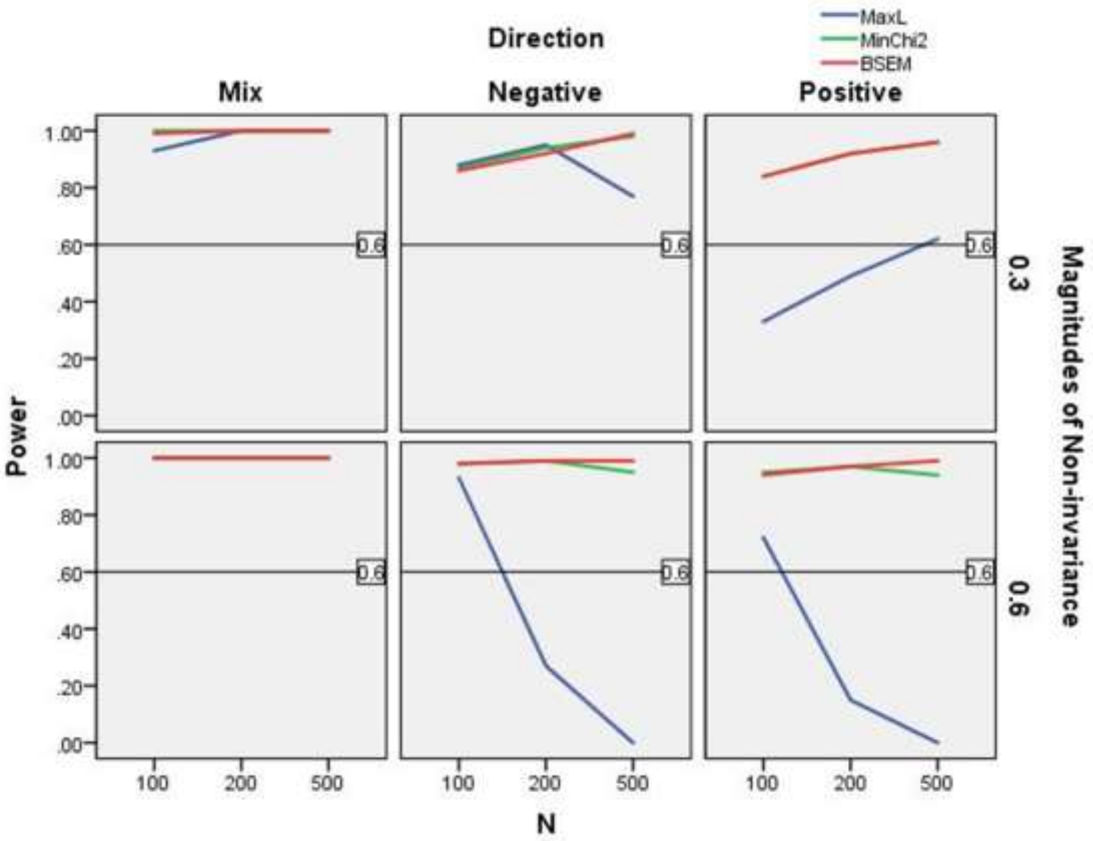
Note: The reference line indicates the power rates as randomly selecting an item as RI.

Figure 1.4: 20% Intercepts Have Non-Invariance



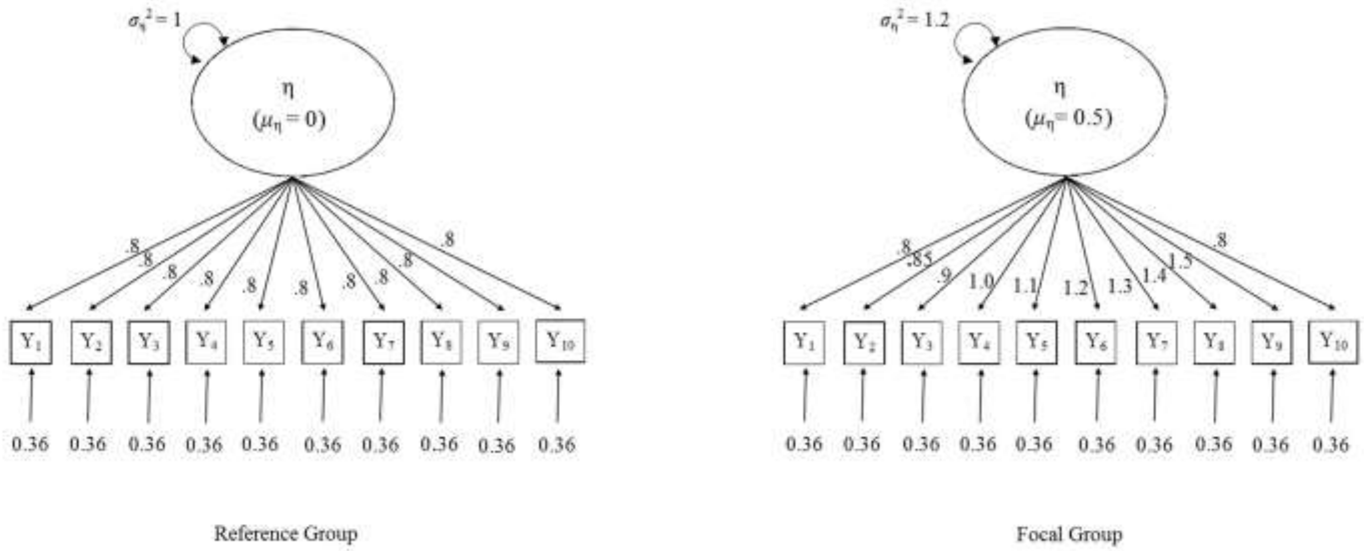
Note: The reference line indicates the power rates as randomly selecting an item as RI.

Figure 1.5: 40% Intercepts Have Non-Invariance



Note: The reference line indicates the power rates as randomly selecting an item as RI.

Figure 2.1: The Generated Non-Invariant Items



Reference Group

Focal Group

Figure 2.2: The Interaction Effect of Bayes Factor Power between Magnitude and Priors

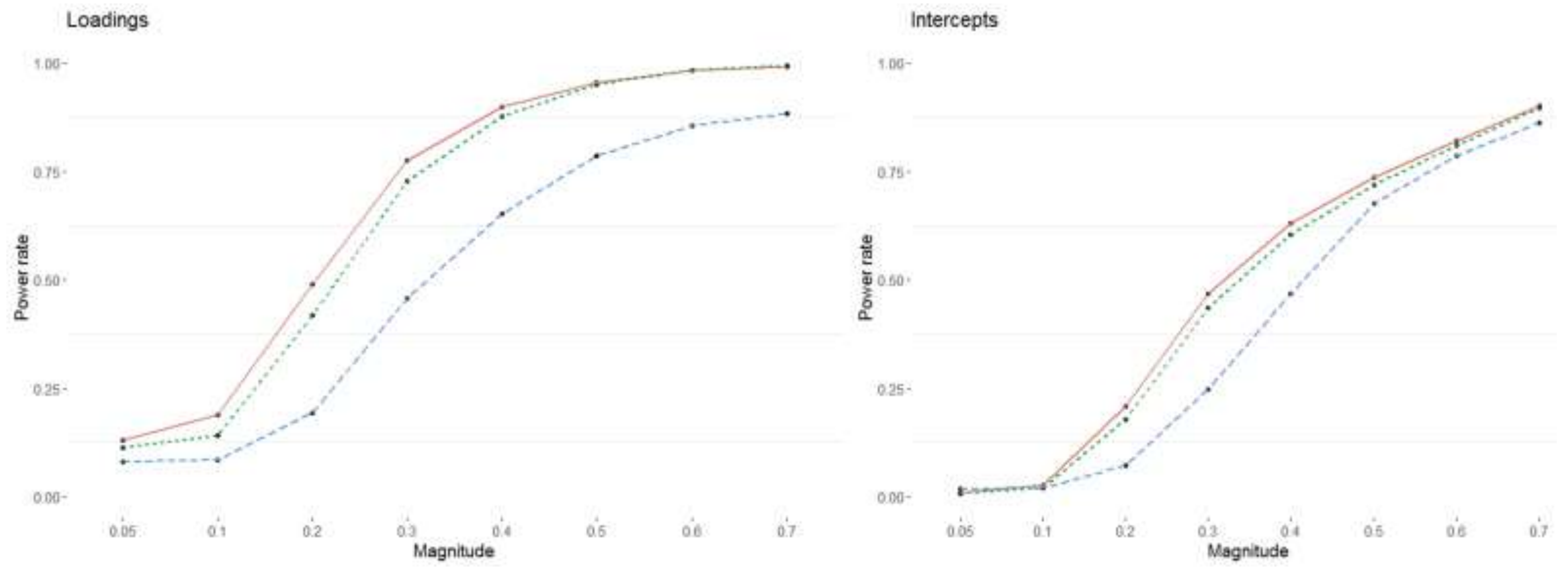


Figure 2.3: The Interaction Effect of Bayes Factor Power between Magnitude and Sample Size

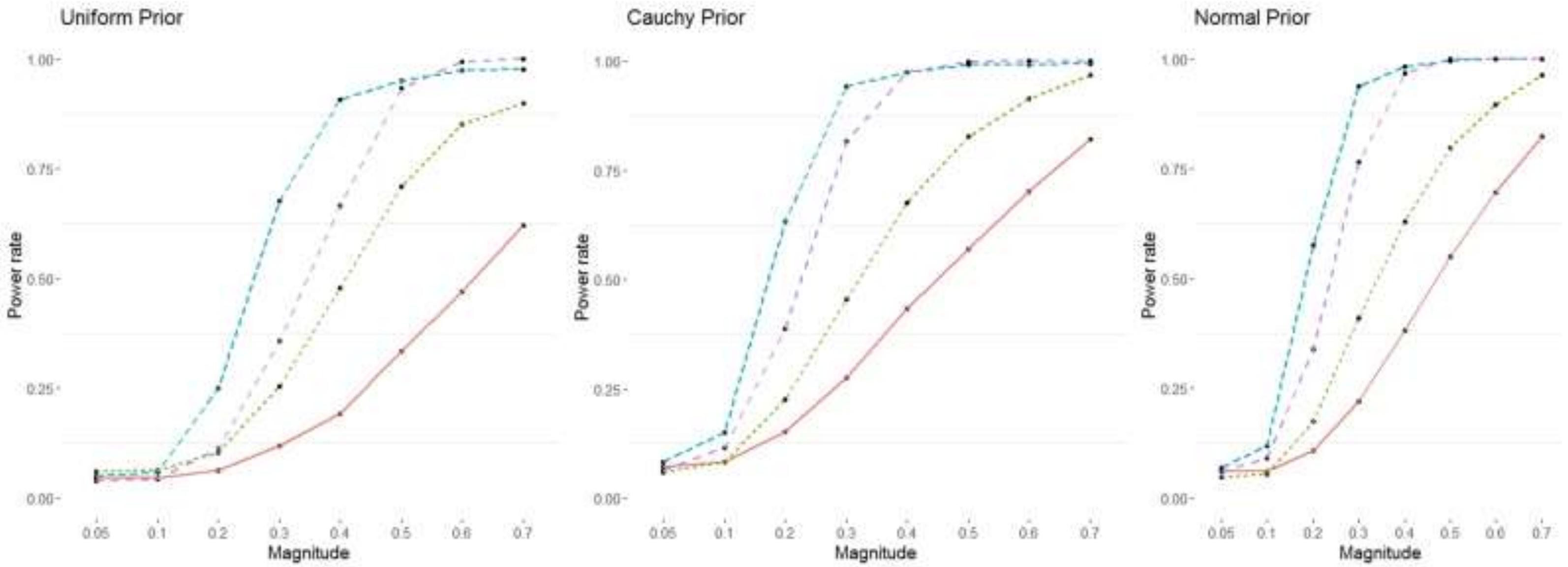


Figure 2.4: The Interaction Effect of ROPE Power between Magnitude and Priors

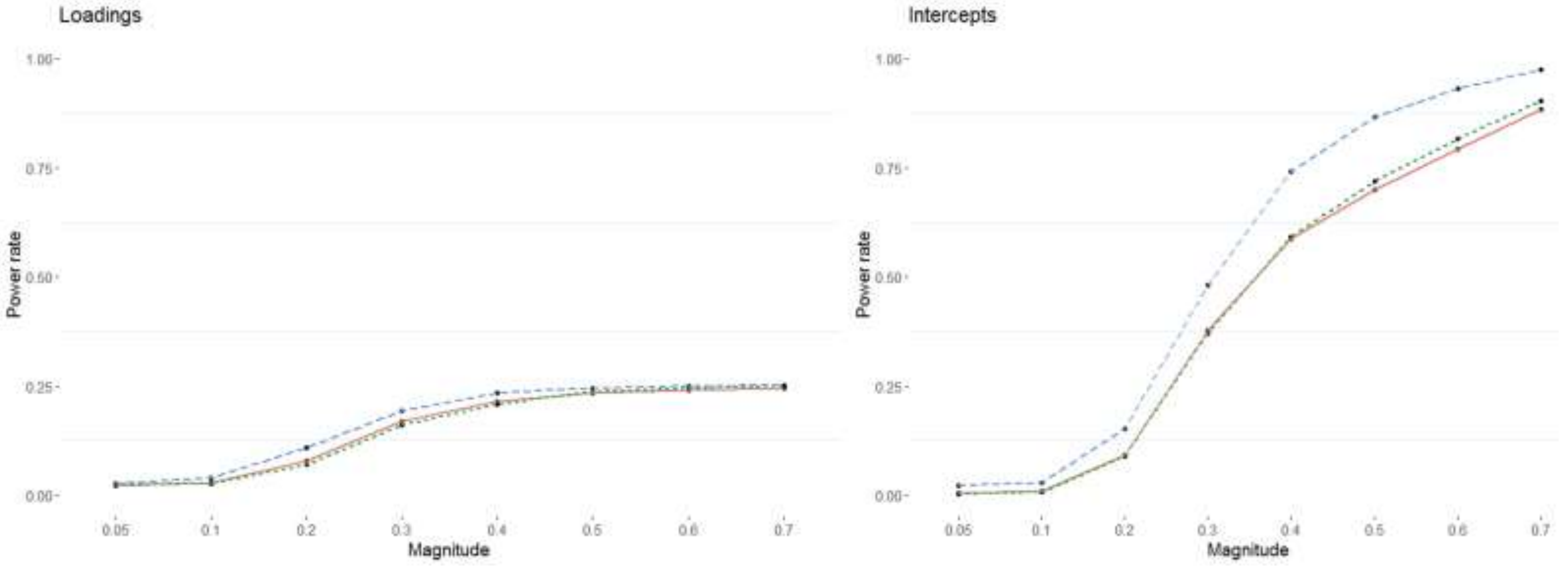


Figure 2.5: The Interaction Effect of Power of ROPE With Zero between Magnitude and Priors

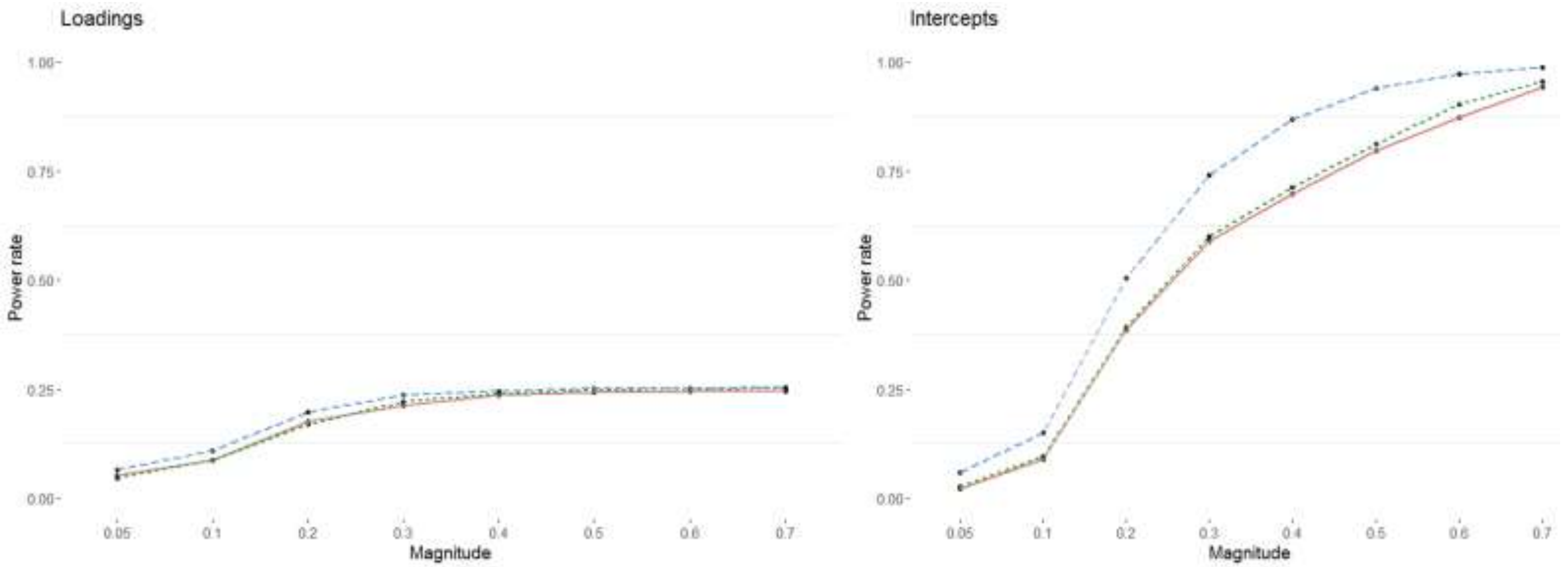


Figure 2.6: The Interaction Effect of ROPE Power between Magnitude and Sample Size

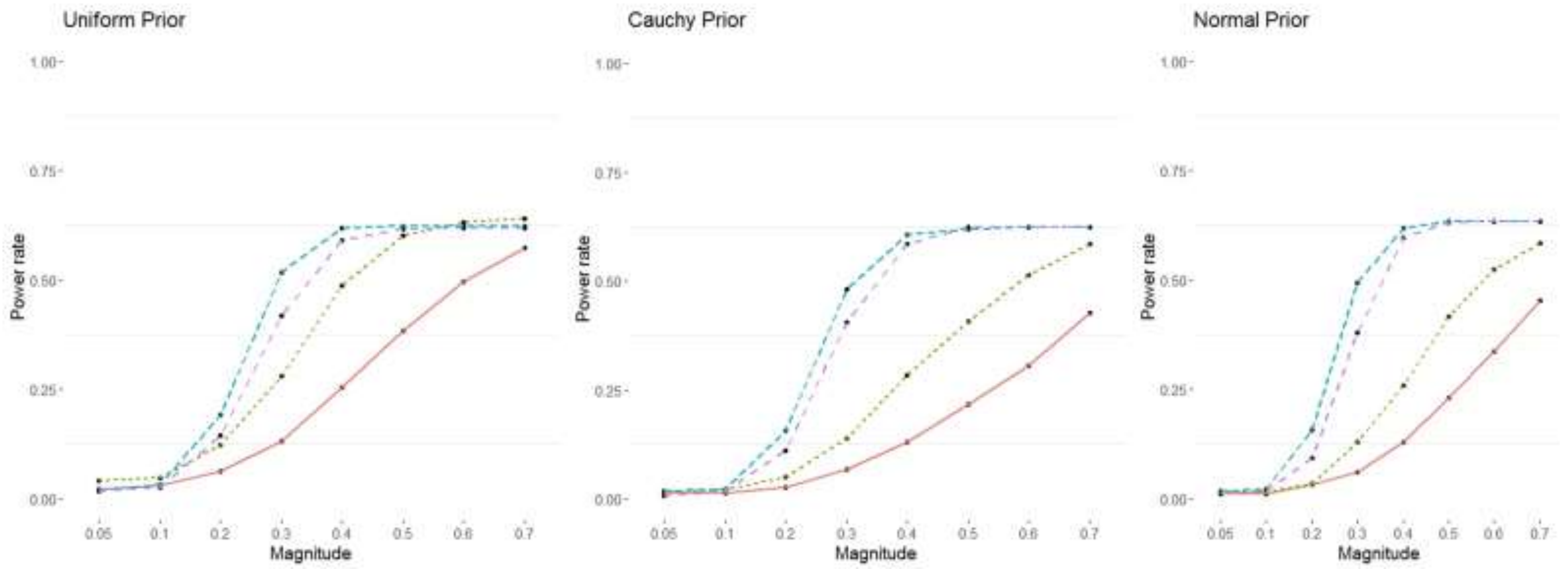


Figure 2.7: The Interaction Effect of Power of ROPE with Zero between Magnitude and Sample Size

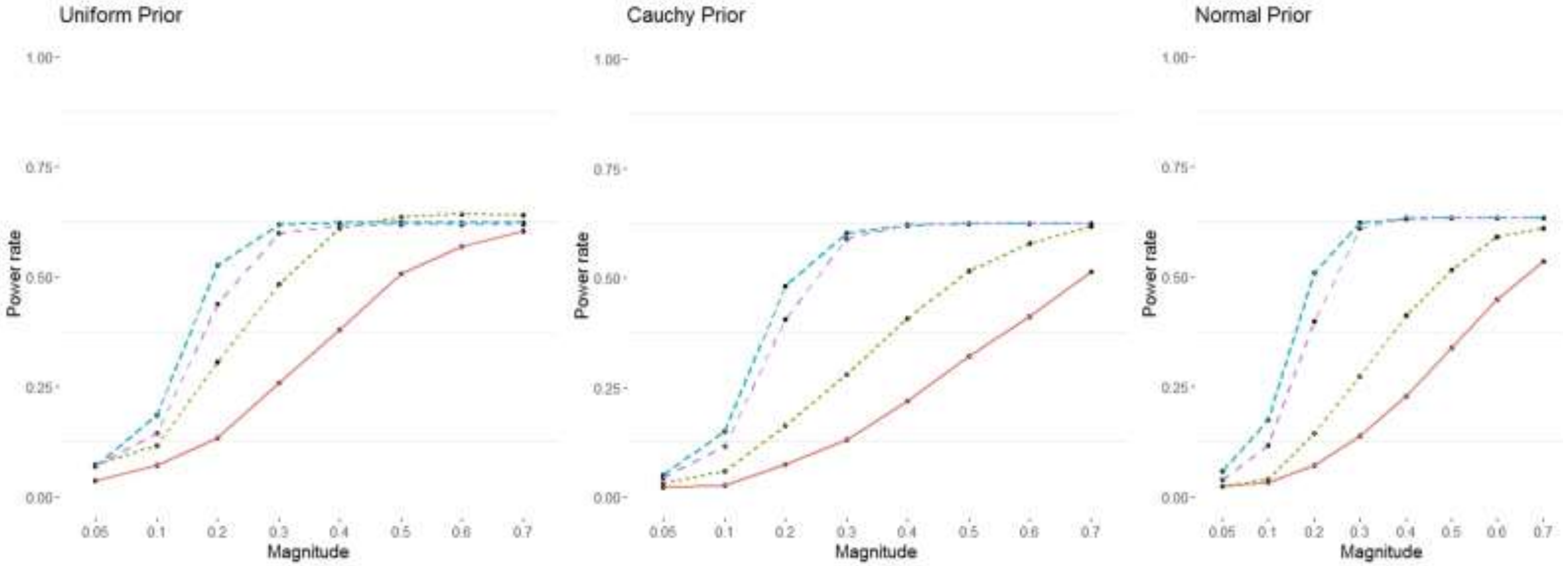


Figure 2.8: The Uncertainty Rate of Bayes Factor

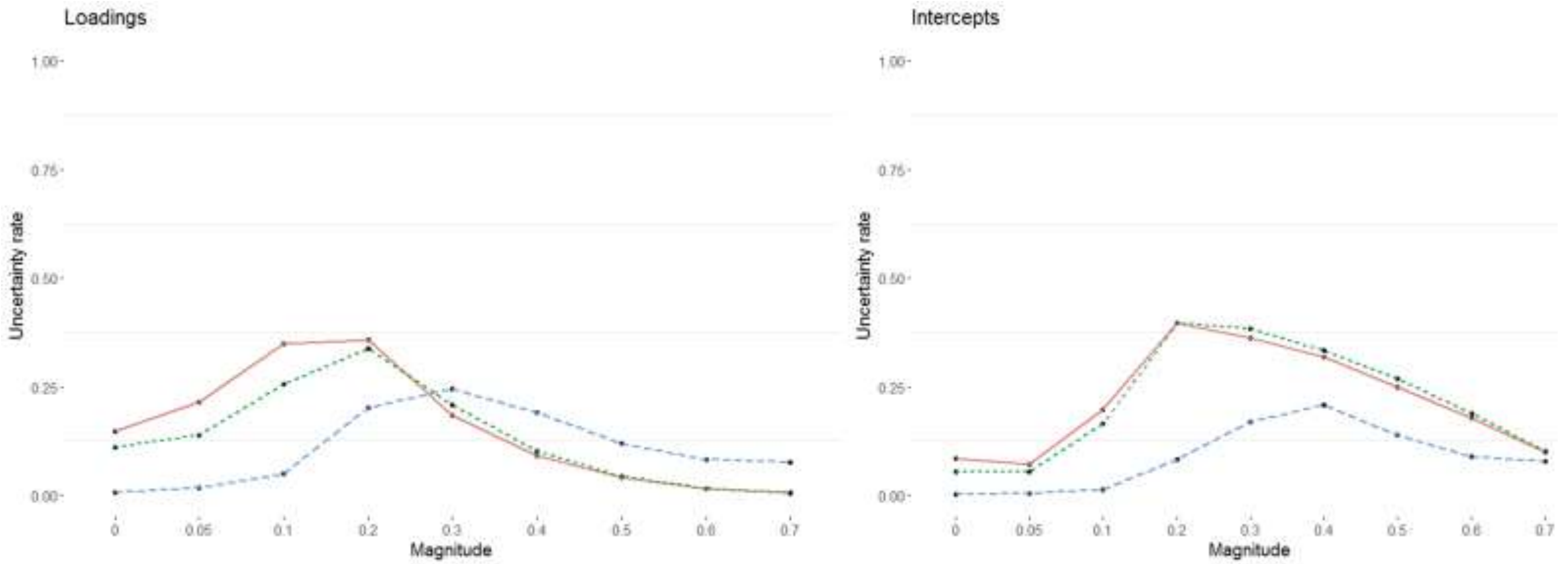


Figure 2.9: The Uncertainty Rate of ROPE

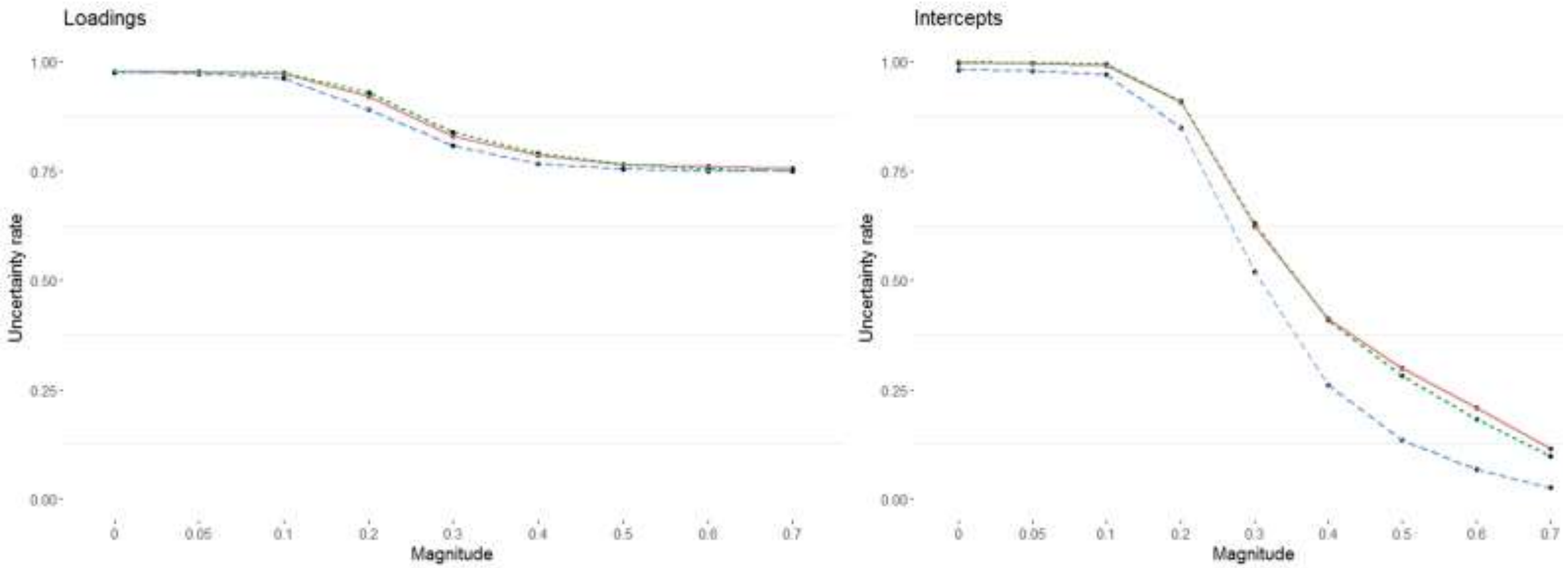


Figure 2.10: The Uncertainty Rate of ROPE with Zero

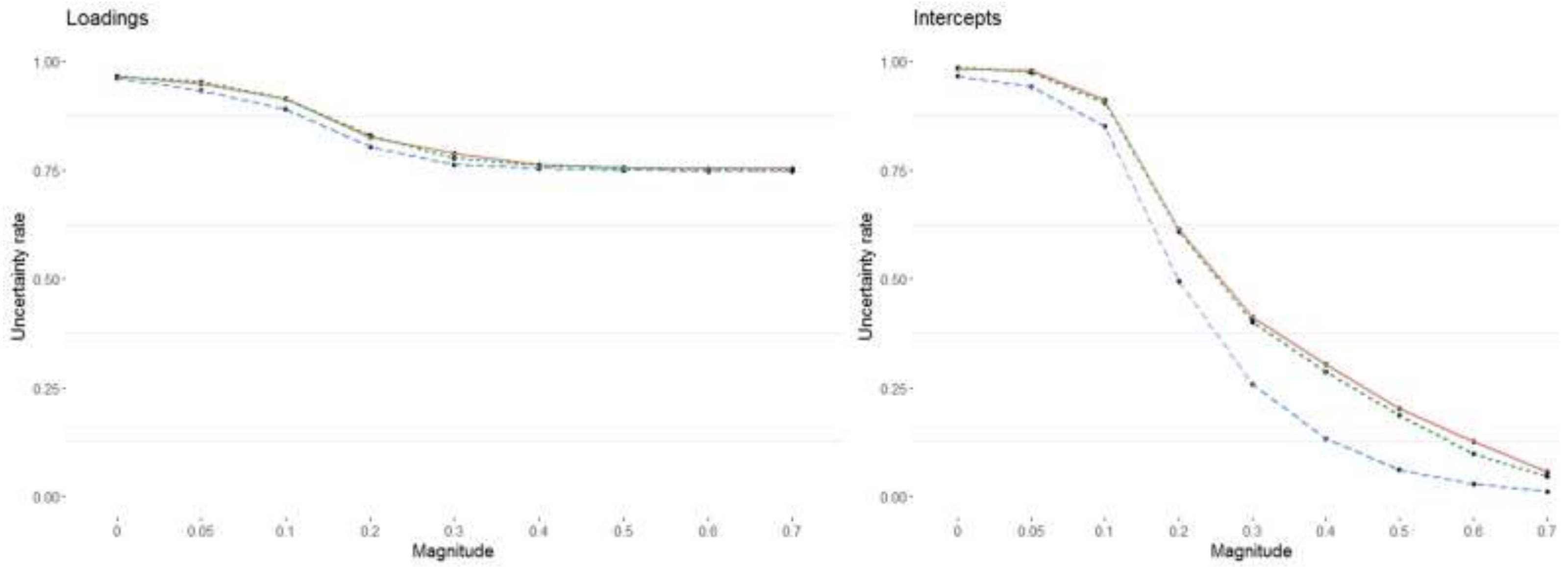


Figure 2.11: Compare the Power Rate of Methods on Loadings and Intercepts

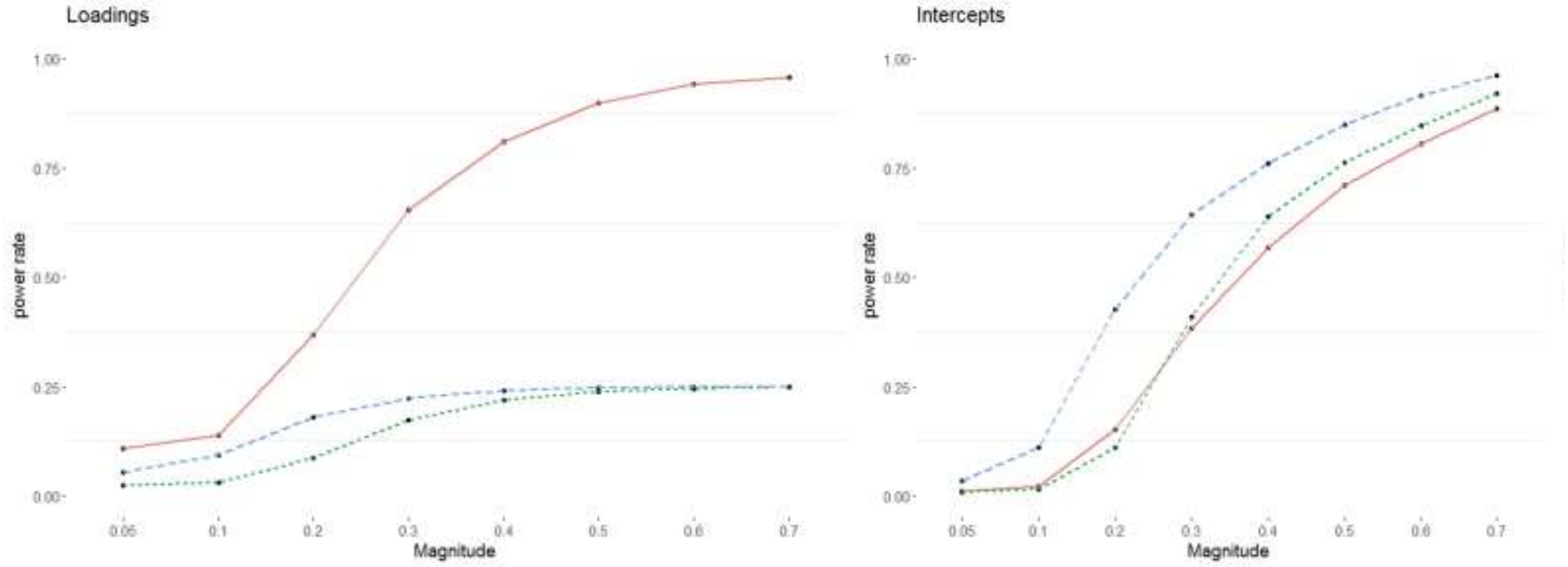


Figure 2.12: Compare the Uncertainty Rate of Methods on Loadings and Intercepts

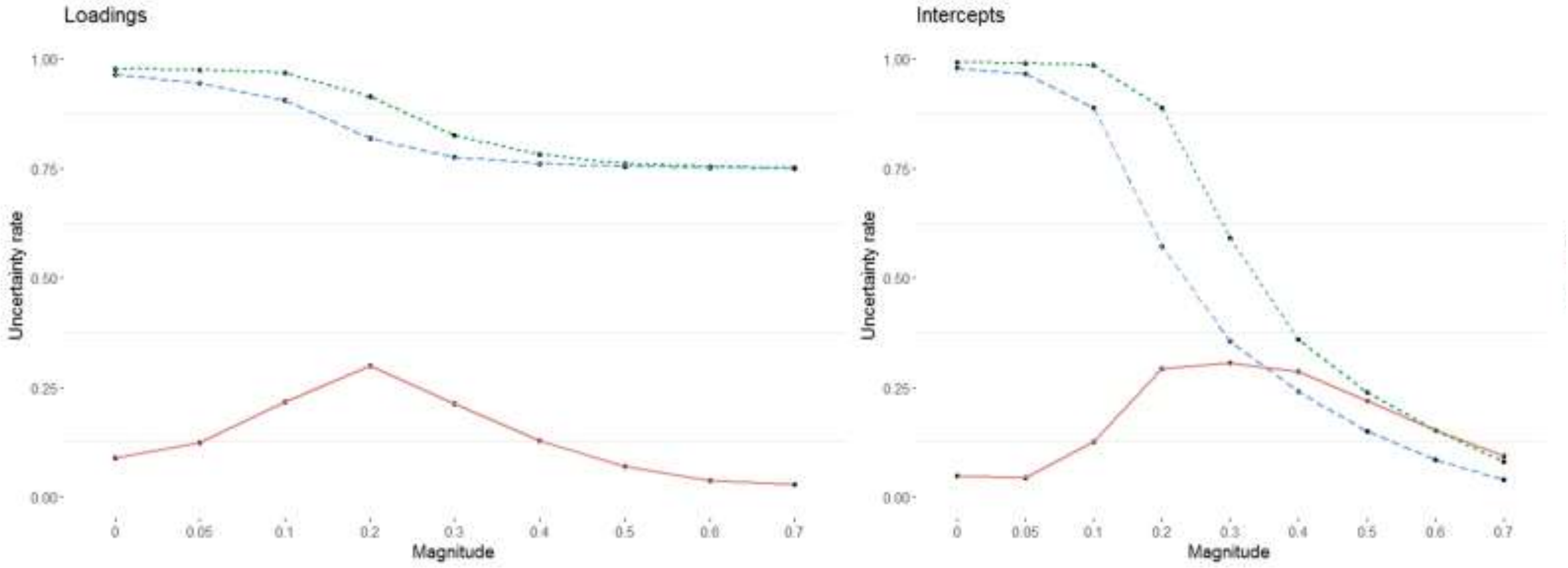


Figure 2.13: Compare the Power Rate of Methods with Uniform, Cauchy and Normal Prior

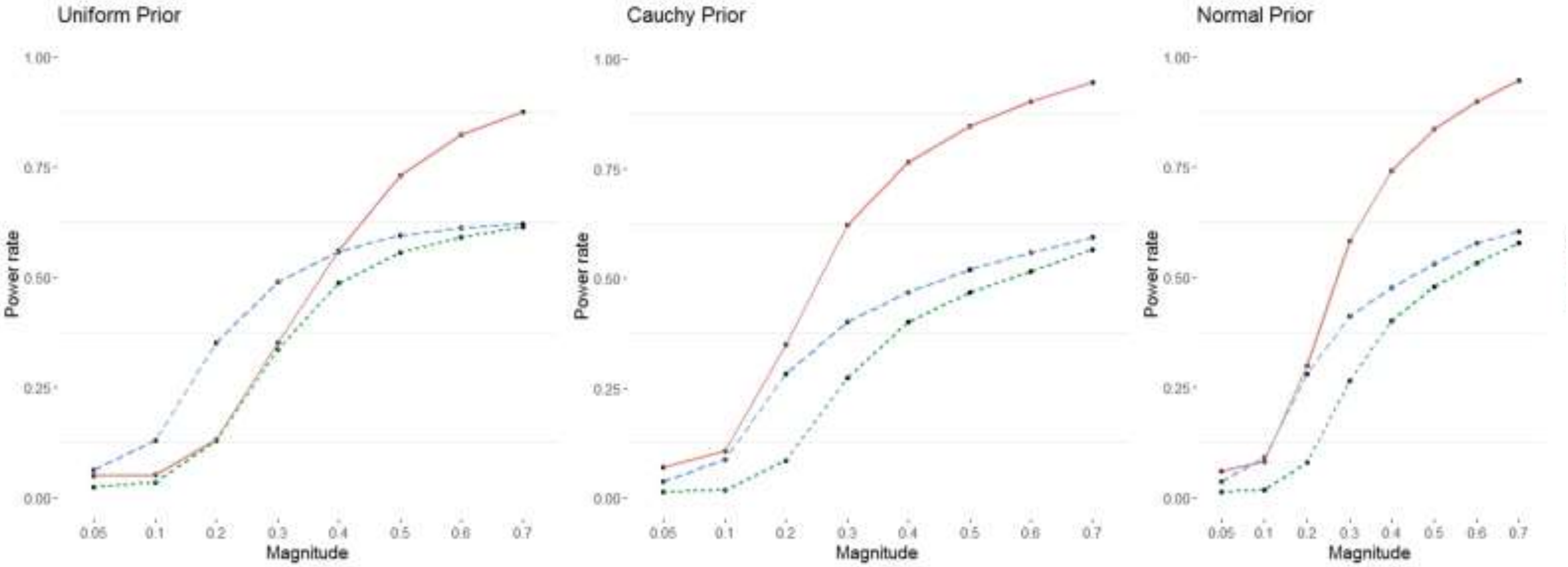


Figure 2.14: Compare the Uncertainty Rate of Methods with Uniform, Cauchy and Normal Prior

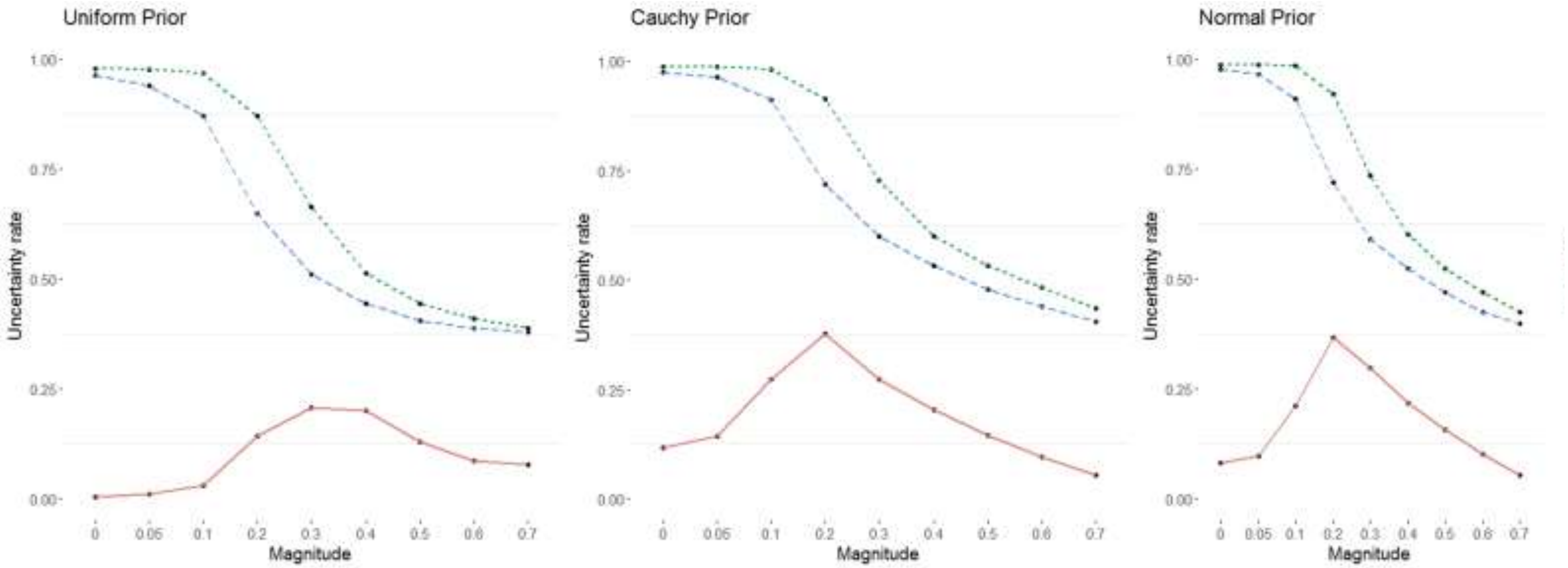


Figure 2.15: Compare the Power Rate of Methods

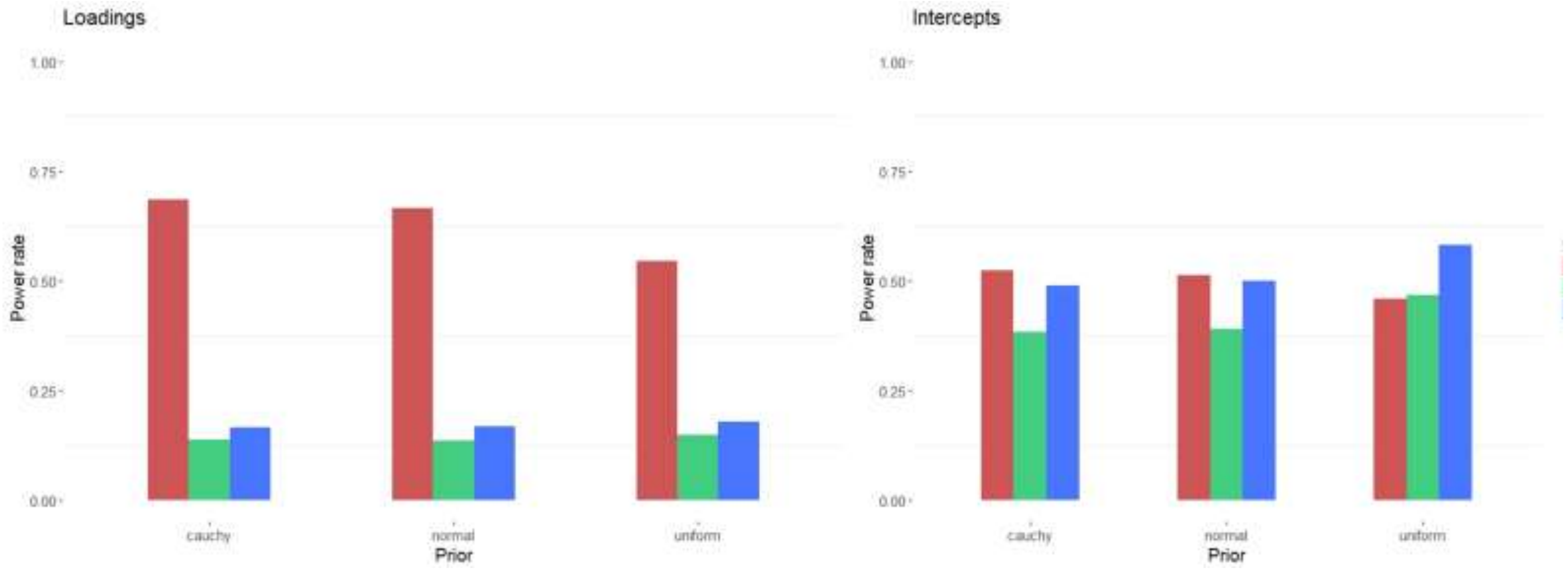


Figure 2.16: Compare the Uncertainty Rate of Methods

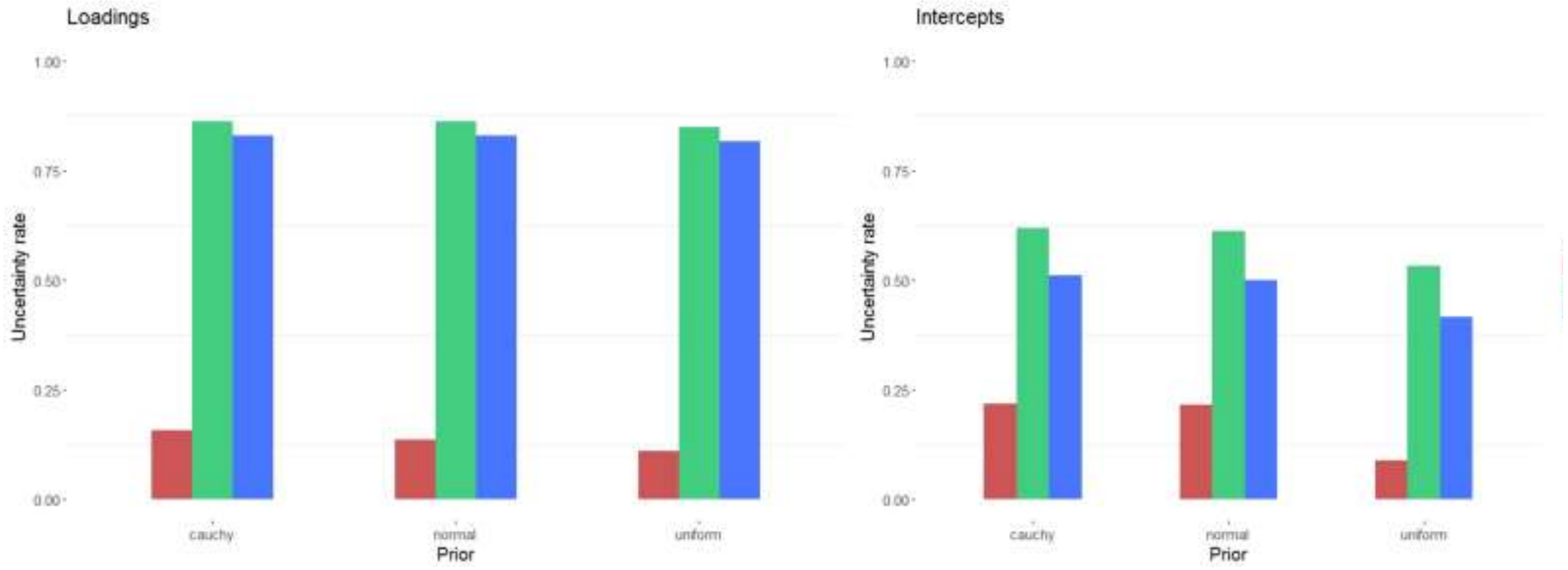
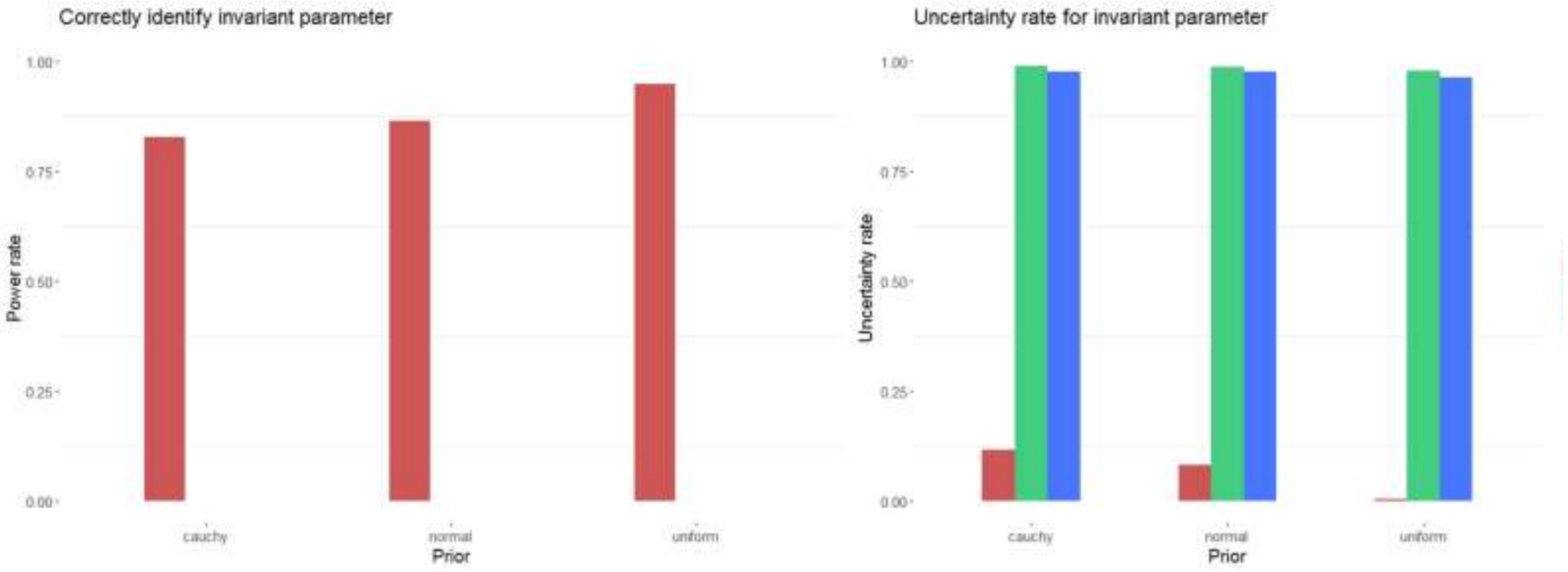


Figure 2.17: Compare the Rate of Correctly Identify Invariance



References

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

Bandalos, D. L. (1997). Assessing sources of error in structural equation models: The effects of sample size, reliability, and model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(3), 177-192.

Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 58-80.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of personality and social psychology*, *100*(3), 407.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588.

Bergh, D. (2015). Sample Size and Chi-Squared Test of Fit—A Comparison Between a Random Sample Approach and a Chi-Square Value Adjustment Method Using Swedish Adolescent Data. In *Pacific Rim Objective Measurement Symposium (PROMS) 2014 Conference Proceedings* (pp. 197-211). Springer, Berlin, Heidelberg.

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

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

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

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

Chorpita, B. F., Yim, L., Moffitt, C., Umemoto, L. A., & Francis, S. E. (2000). Assessment of symptoms of DSM-IV anxiety and depression in children: A revised child anxiety and depression scale. *Behaviour research and therapy*, *38*(8), 835-855.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214-226.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 204-223.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on?. *Perspectives on Psychological Science*, *6*(3), 274-290.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*, 781.

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

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398-409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari A. & Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Geman, S., & Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in Computer Vision* (pp. 564-584).

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, *41*(2), 421-440.

Gollwitzer, M., & Melzer, A. (2012). Macbeth and the joystick: Evidence for moral cleansing after playing a violent video game. *Journal of Experimental Social Psychology*, *48*(6), 1356-1360.

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*(2), 306-324.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological methods, 16*(3), 319-336.

Hoijtink, H., Béland, S., & Vermeulen, J. A. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological methods*, *19*(1), 21.

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.

Jeffreys, H. (1935, April). Some tests of significance, treated by the theory of probability. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 31, No. 2, pp. 203-222). Cambridge University Press.

Jeffreys, H. (1961). *The theory of probability*. OUP Oxford.

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

Jöreskog, K. G. (1971). Simultaneous Factor Analysis in Several Populations. *Psychometrika*, *36*(4), 49-426.

Jung, E., & Yoon, M. (2016). Comparisons of three empirical methods for partial factorial invariance: forward, backward, and factor-ratio tests. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 567-584.

Jung, E., & Yoon, M. (2017). Two-step approach to partial factorial invariance: Selecting a reference variable and identifying the source of noninvariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(1), 65-79.

Kary, A., Taylor, R., & Donkin, C. (2016). Using Bayes factors to test the predictions of models: A case study in visual working memory. *Journal of Mathematical Psychology*, *72*, 210-219.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Klugkist, I., van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know?. *International Journal of Behavioral Development*, *35*(6), 550-560.

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.

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.

Konijn, E. A., van de Schoot, R., Winter, S. D., & Ferguson, C. J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. *Communication Methods and Measures*, *9*(4), 280-302.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299-312.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178-206.

Langer, M. M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Doctoral dissertation, The University of North Carolina at Chapel Hill).

Lantz, B. (2013). The large sample size fallacy. *Scandinavian journal of caring sciences*, *27*(2), 487-492.

Liao, X. (2012). *The impact of measurement non-equivalence on second-order latent growth curve modeling* (Doctoral dissertation, University of Oklahoma).

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362-375.

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*(4), 251-265.

Ly, A., Verhagen, J., & Wagenmakers, E. J. (2016). Harold Jeffreys’s default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19-32.

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

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E. J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*(1), e1.

McNeish, D., An, J., & Hancock, G. R. (2018). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of personality assessment*, *100*(1), 43-52.

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

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

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology*, *97*(5), 1016.

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

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

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

Millsap, R. E. (2012). Statistical approaches to measurement invariance. New York, NY: Routledge.

Mou, Y., Berteletti, I., & Hyde, D. C. (2018). What counts in preschool number knowledge? A Bayes factor analytic approach toward theoretical model development. *Journal of experimental child psychology, 166,* 116-133.

Mulder, J., & Wagenmakers, E. J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*, *72*, 1-5.

Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. *Mplus Web Notes*, *17,* 1-48.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79-95.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, *47*(1), 90-100.

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

Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. *A handbook for data analysis in the behavioral sciences: Methodological issues*, 199-228.

Shi, D., Song, H., & Lewis, M. D. (2017). The impact of partial factorial invariance on cross-group comparisons. *Assessment*, 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.

Shi, D., Maydeu-Olivares, A. & Distefano, C. (2018). The Relationship between the Standardized Root Mean Square Residual and Model Misspecification in Factor Analysis Models. *Multivariate Behavioral Research.* 10.1080/00273171.2018.1476221

Shi, D., Lee. T., & Maydeu-Olivares (2018). Understanding the Model Size Effect on SEM Fit Indices. *Educational and Psychological Measurement*.

Shi, D., Song, H., Distefano, C. Maydeu-Olivares, A., McDaniel, H & Jiang, Z. (2018). Evaluating Factorial Invariance: An Interval Estimation Approach using Bayesian Structural Equation Modeling (BSEM). *Manuscript submitted for publicaiton*.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, 3-23.

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.

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

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*(3), 748-761.

Steinmetz, H. (2011). Estimation and comparison of latent means across cultures. Cross-cultural analysis: Methods and applications, 85-116.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American psychologist*, *44*(10), 1276.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic bulletin & review*, *25*(1), 102-113.

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.

Van Den Hout, M., Gangemi, A., Mancini, F., Engelhard, I. M., Rijkeboer, M. M., Van Dams, M., & Klugkist, I. (2014). Behavior as information about threat in anxiety disorders: A comparison of patients with anxiety disorders and non-anxious controls. *Journal of behavior therapy and experimental psychiatry*, *45*(4), 489-495.

Van Den Hout, M., Gangemi, A., Mancini, F., Engelhard, I. M., Rijkeboer, M. M., van Dam, M., & Klugkist, I. (2017). “Behavior as information about threat in anxiety disorders: A comparison of patients with anxiety disorders and non-anxious controls”: Erratum.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic bulletin & review*, *25*(1), 143-154.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*(2), 217.

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of mathematical psychology*, *72*, 171-182.

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology*, *60*(3), 158-189.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: The case of psi. Journal of Personality *and Social Psychology*, *100*(3), 426-432.

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... & Matzke, D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic bulletin & review*, *25*(1), 35-57.

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

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

Williams, M., Bååth, R. A., & Philipp, M. C. (2017). The Bayes factor: A bridge to Bayesian inference.

Wong, T. M., & Van de Schoot, R. (2012). The effect of offenders' sex on reporting crimes to the police. *Journal of interpersonal violence*, *27*(7), 1276-1292.

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *The Journal of Experimental Education*, *80*(1), 26-44.

Whisman, M. A., & Judd, C. M. (2016). A cross-national analysis of measurement invariance of the Satisfaction with Life Scale. *Psychological Assessment, 28*(2), 239.

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42-57.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*(3), 532-547.

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

Appendices

Appendix A: Footnotes

1. Alternatively, one can begin such tests by fitting a model with all the parameters constrained to be equal, and then progressively relaxing certain equality constraints. Further information on this approach can be found in Stark, Chernyshenko, and Drasgow (2006), Yoon and Milsap (2007), 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. Woods (2009) ranked order the items based on their LR/Δ*df*. In our study, we used LR instead of ratio of LR/Δ*df*, because Δ*df* (=2) was constant across all conditions.
3. Please pay attention to the notation of BF. $BF_{10} > 3$ indicates the data is in favor of alternative H_1.
4. Within the same range, the power rate of BF were generally lower than Wald at $p<0.05$ level in most of conditions.
5. We followed the way by Wagenmakers et al (2018) to interpret the results from JASP. All the models take the power rate as outcome. Data has been divided to two parts based on the location of non-invariance.
6. The only agreement that Both BF and Bayesian estimation got was on Item 10.
7. The concept of shrinkage estimation originally from hierarchical models.

Appendix B: Supplement Figures

Figure S1: The Interaction Effect of Methods and Directions at Levels of Percentage

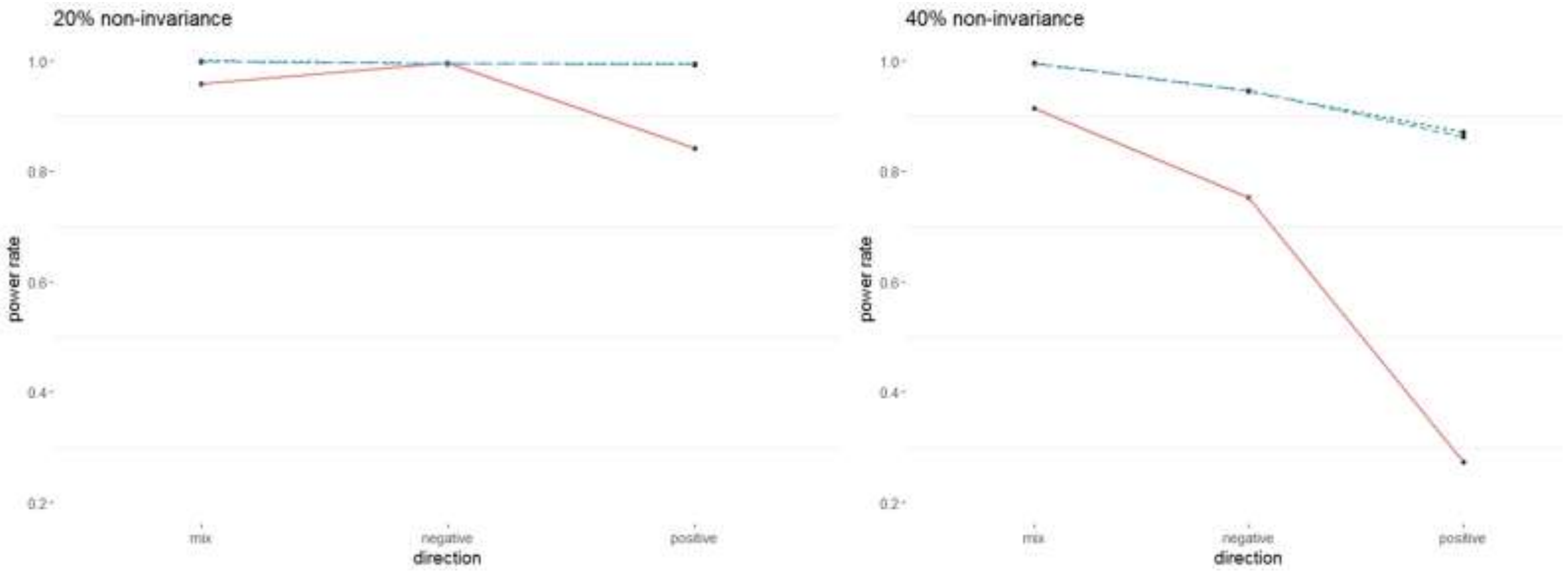


Figure S2: The Interaction Effect of Methods and Directions at Levels of Sample Size

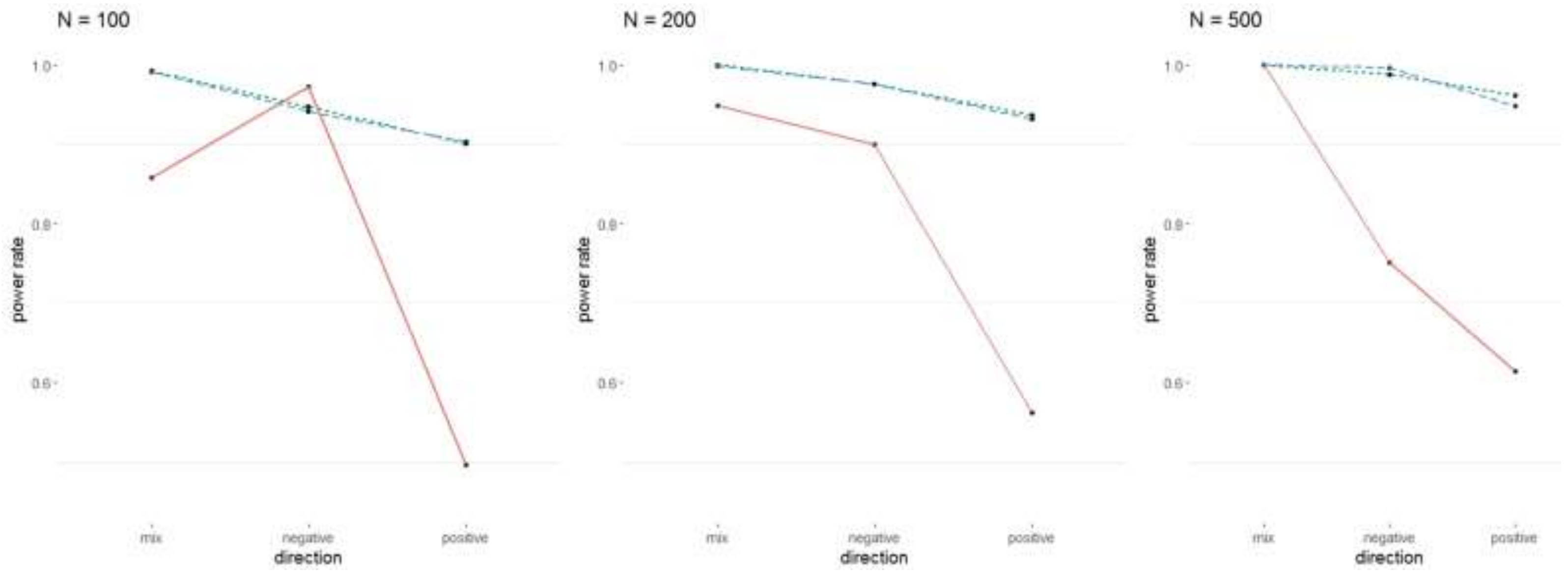


Figure S3: The Interaction Effect of Methods and Directions at Levels of Magnitudes

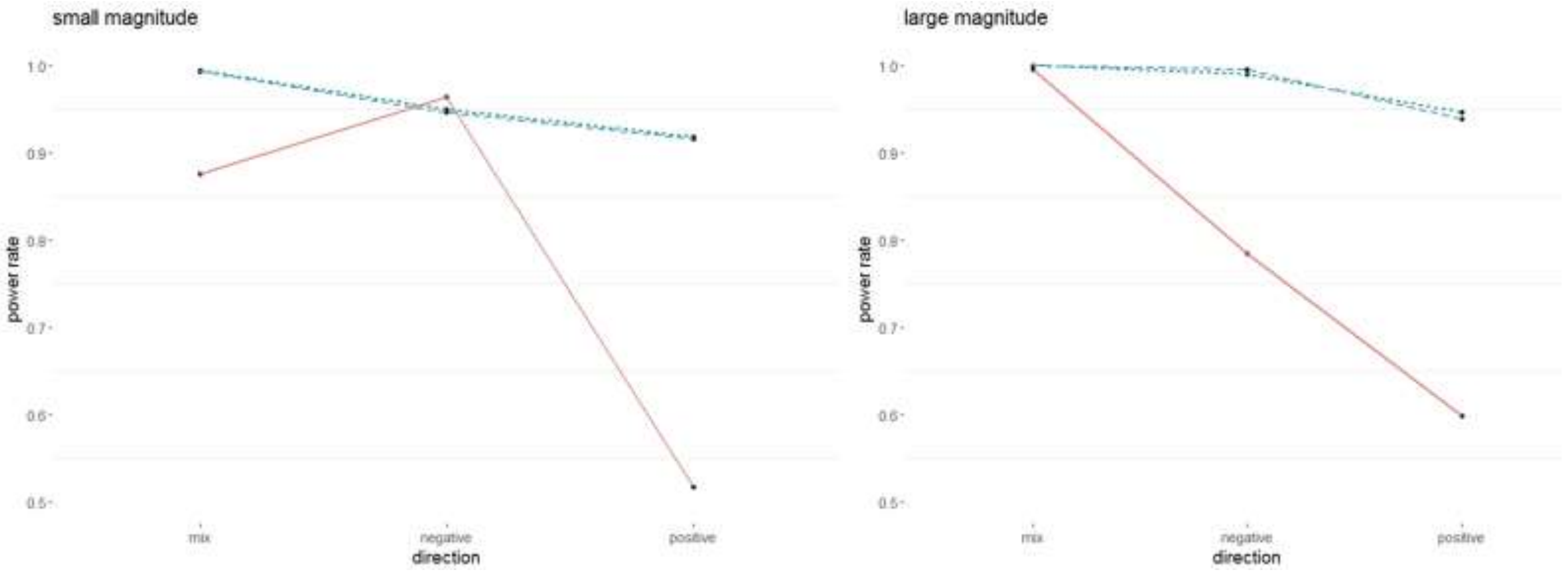


Figure S4: The Interaction Effect of Methods and Directions at Levels of Locations

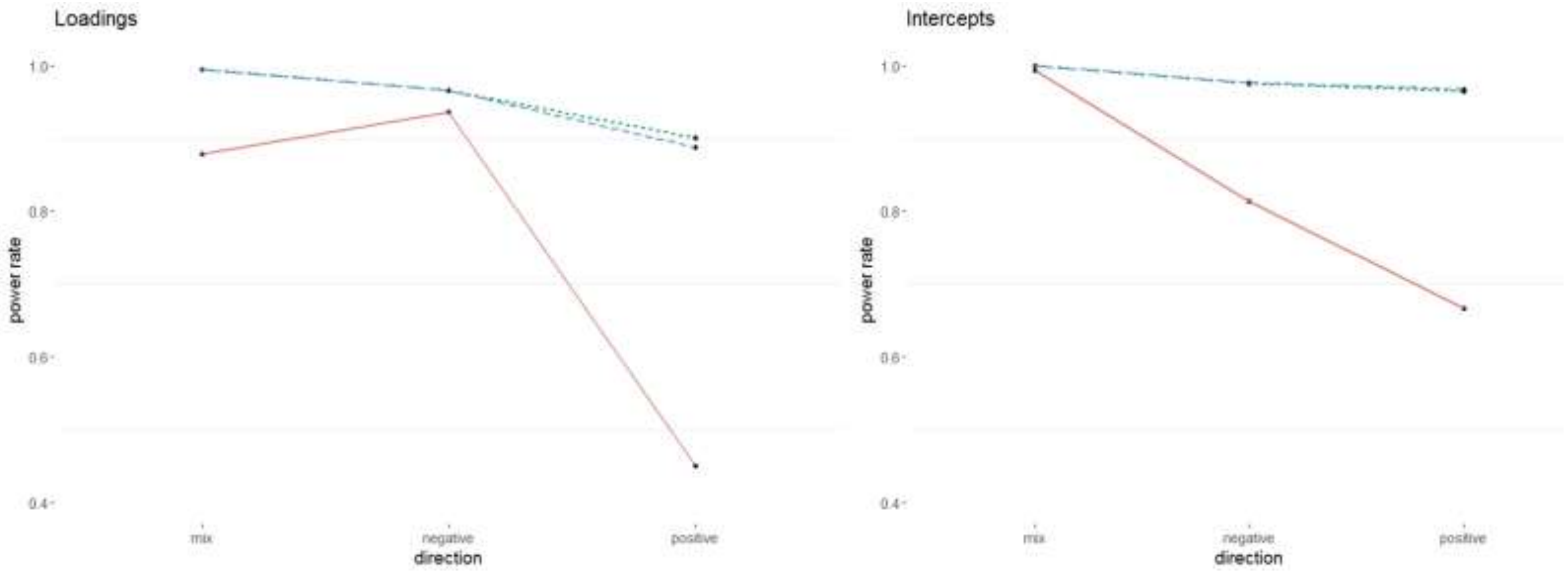


Figure S5: The Interaction Effect of Methods and Magnitude at Levels of Percentage

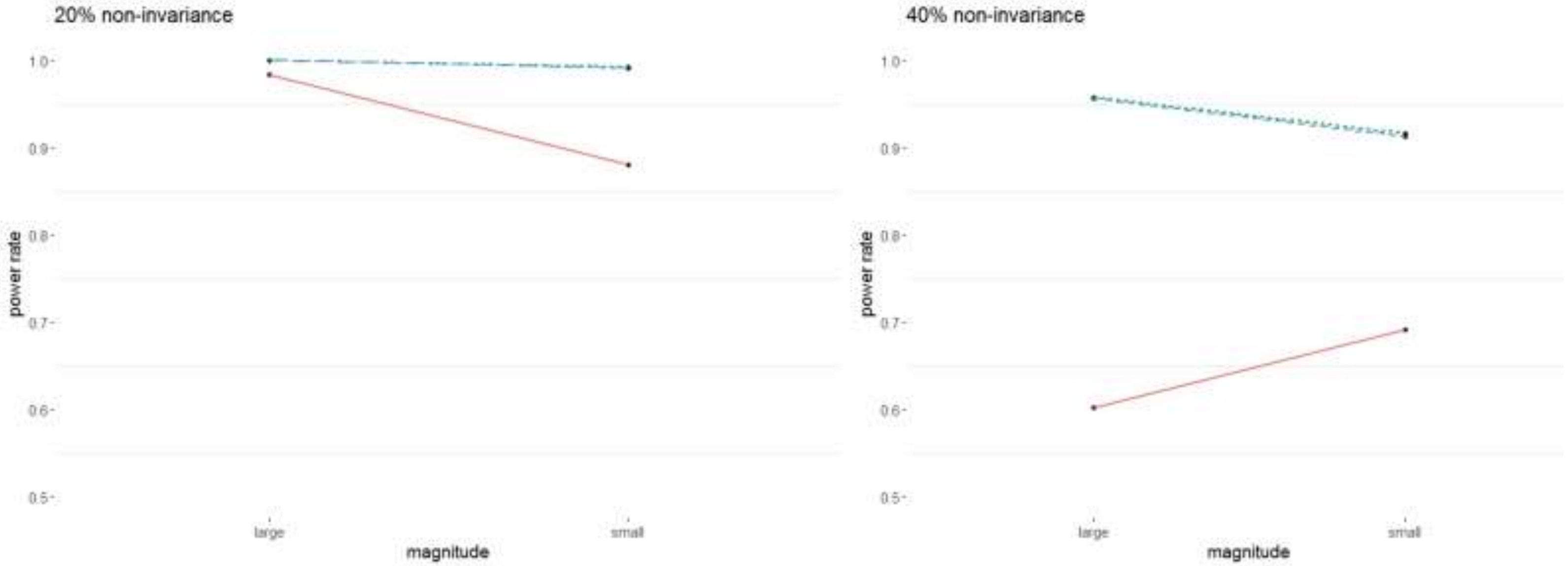


Figure S6: The Interaction Effect of Methods and Magnitude at Levels of Sample Size

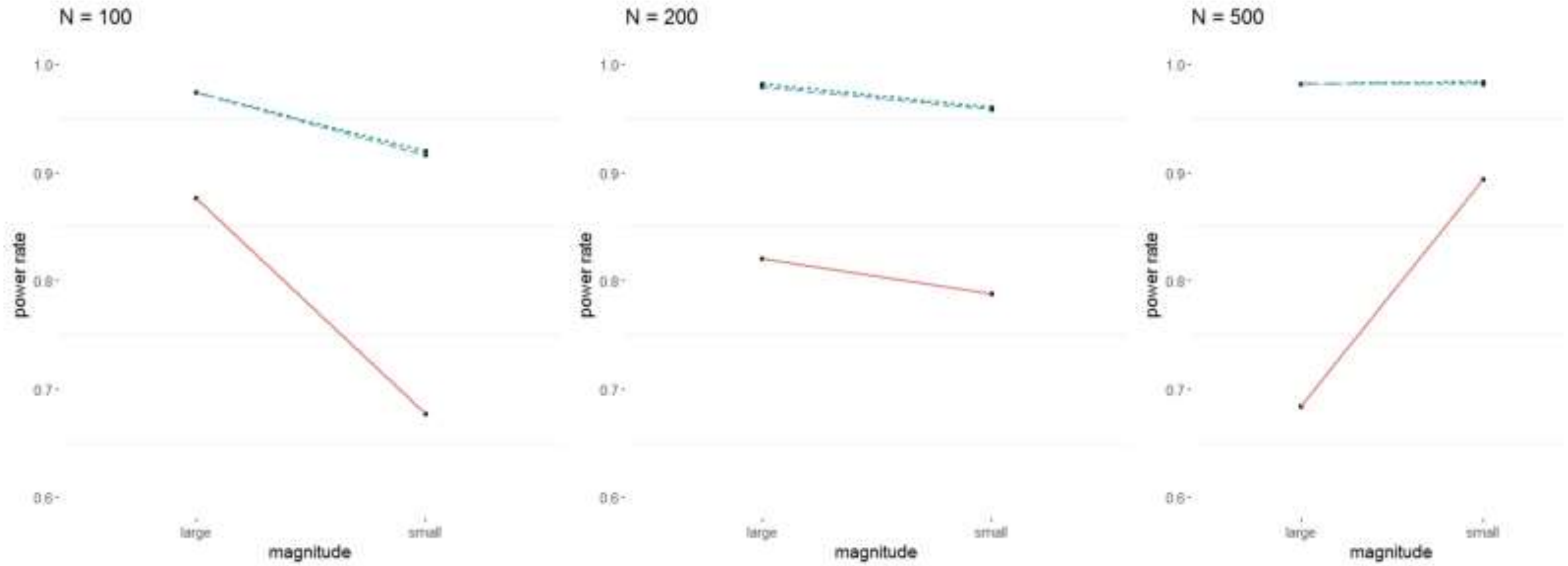


Figure S7: The Interaction Effect of Methods and Magnitude at Levels of Directions

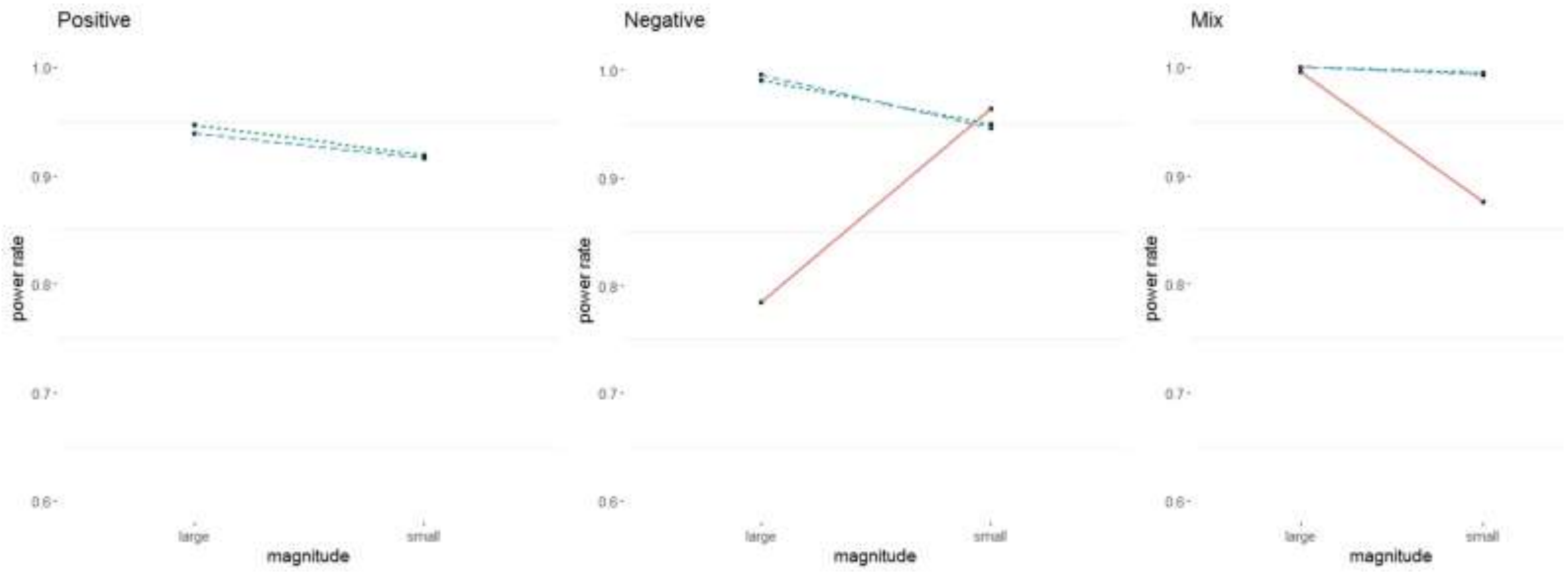
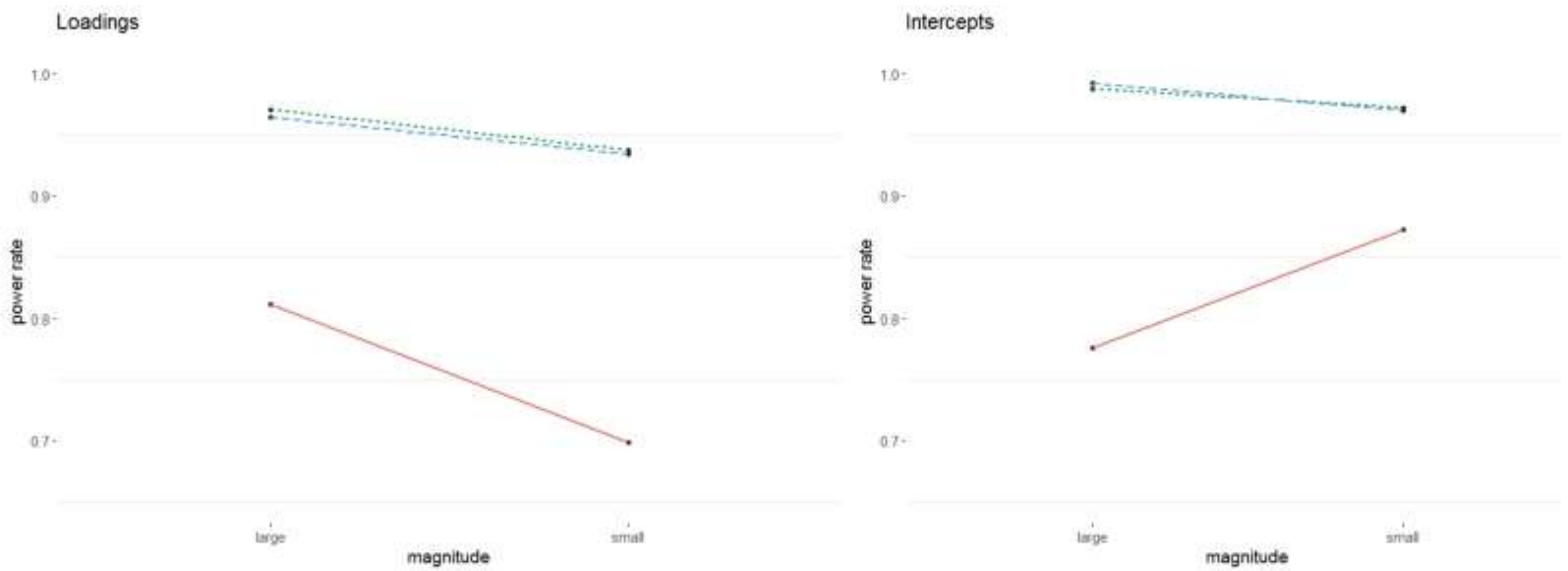


Figure S8: The Interaction Effect of Methods and Magnitude at Levels of Locations


Loadings
Intercepts
power rate
large
small
magnitude