A Bayesian Approach: Measurement Invariance Testing and Prediction

By

RUI JIANG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSPPHY

in

Quantitative Psychology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Philippe Rast

Siwei Liu

Mijke Rhemtulla

Committee in Charge

2022

Table of Contents

List of Figures

List of Table

Abstract

Measurement invariance (MI) is an important assumption in testing group-mean differences. MI describes the conditions where the measurement model of a latent construct is equivalent across different groups. Despite the importance of establishing full MI, it is unrealistic to fulfill all levels of MI (i.e., pattern, weak, strong, strict) in the real world. Alternatively, researchers may choose to fit partially invariant models. Yet, this popular approach requires the exact identification of non-invariance and leaves the question open about which model fitting strategies should be selected. Thus, the current dissertation seeks to answer two related questions via simulation studies: (1) How to use sparse Bayesian estimation methods (i.e., the Horseshoe prior) for detecting non-invariant items; (2) Can we improve the predictive performance of partially invariant models using the horseshoe (HS) prior?

The first study aims to demonstrate how to use the HS prior for detecting non-invariant items. I discuss how to approach the identification of non-invariant items as a variable selection problem and I describe how to identify non-invariant items within a Bayesian framework relying on the HS prior used in Bayesian model selection. A simulation study is introduced to investigate the performance of the HS priors in identifying non-invariant items under various conditions. The simulation conditions include sample size, parameter difference, scale length, and item reliability. The results showed that the HS prior approach almost always accurately identified non-invariant items. A large sample size and a high item-reliability can facilitate the identification of non-invariant items even when the amount of non-invariance was small. For identifying invariant items, the HS prior approach exhibited an almost perfect performance.

The second study seeks to improve the predictive performance of partially invariant models. Common model solutions for partially invariant models usually focus on explaining

the underlying mechanism of psychological phenomena. Current approaches may not generalize well to new and unseen data since MI is only considering the characteristics of the current sample at hand. Recently, the field has seen an increase in the interest in evaluating psychological assessments based on out-of-sample performance. Here, I evaluate an out-of-sample prediction-focused strategy based on partially invariant items. I employed a HS prior model to mimic the idea of Bayesian-Model-Averaging for improving the predictive performance of partially invariant models. The results of a simulation study indicated that the HS prior model outperformed the other commonly used model fitting strategies (i.e., fully constrained model, partially constrained model, freely estimated model) under most conditions. Sample size, parameter differences and item-reliability showed differential impacts on models' predictive performance.

The third study illustrated with an example how to use the HS prior approach for empirical analyses. First, the DERS-9 scale was assessed for item-level MI between genders and between two measurement occasions with a sample of 300 for each group. Next, a partially invariant SEM model was fitted with a sample of 728 college students, where the partially invariant status of the self-esteem scale between genders was confirmed, and then the HS prior model and the freely estimated model were fitted for comparison where peer victimization was regressed on self-esteem between genders. The results indicated that the HS prior performed well in terms of predictions with empirical data.

In conclusion, this dissertation discussed and explored potential solutions of two major issues with measurement invariance: non-invariant item detections, and the predictive performance of partially invariant models. The results of two simulation studies indicated that the HS prior approach is a viable alternative to traditional methods for identifying non-invariant items and fitting partially invariant models. Finally, the implications and limitations of this set of studies, along with recommendations for future studies were discussed.

1. Chapter 1

Historical Background of Measurement Invariance

1.1 Goals of Measurement Invariance Assessment

Measurement invariance (MI) is an important assumption in testing group-mean differences (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 2002). MI describes the conditions under which the measurement model of a latent construct is equivalent across heterogeneous populations, different measurement occasions, or survey formats (French & Finch, 2016; Millsap & Kwok, 2004; Millsap & Meredith, 2007). Essentially, MI assesses the invariance of factor structures over a given number of group variables such as measurement locations, time points, assessment formats, and populations' characteristics (e.g., gender, age group) (e.g., Millsap, 2011). Ideally, the difference in factor scores should only be a function of the true difference in measured attributes, but not a function of these grouping variables. Otherwise, the group comparison may become theoretically uninterpretable (can we compare apples with bananas?), and statistically biased. Therefore, the establishment of MI ensures the meaningful comparisons of factor scores across different groups, the selection and diagnostic accuracy with different populations, and the validity of the parameter estimates in structural models (French & Finch, 2016; Millsap & Kwok, 2004; Millsap & Meredith, 2007; Shi, Song & Lewis, 2019). In contrast, the violation of MI may produce inconsistent results, artificial differences, or equivalences in test scores between groups (Byrne, Shavelson, & Muthén, 1989; Cheung & Rensvold, 2002; Johnson, Measde & DuVernet, 2009). For example, in a Driver Knowledge Test, an individual with a fixed level of driving knowledge should produce the same test result using an invariant test, regardless of the location where the test is taken (e.g., driving school versus DMV). With a non-invariant math test, male students may

consistently produce higher scores than female, even though they are on the same level of mathematical competence. MI assessment plays a key role in various areas of research and decision-making stages.

Multigroup Confirmatory Factor Analysis (MGCFA) is the most commonly used method in assessing MI at model/scale level (Jöreskog, 1971; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2000; Zumbo & Koh, 2005). The estimation of measurements in CFA involves four components: measurement structure, factor loading, intercepts and residual variance (Millsap, 2011). Each of these four components is associated with four levels of invariance with increasing restrictions. The first level of invariance is configural invariance which concerns the group equality in the measurement structure. Achieving this level of invariance ensure the numbers of constructs and the numbers of items related to each construct is the same between groups. The second level of invariance is weak invariance which concerns the group equality in the measurement structure and factor loadings. Achieving this level of invariance ensures the quantitative association between all scale items to the latent construct is the same between groups. The third level of invariance is strong invariance which concerns the group equality in measurement structure, factor loadings as well as intercepts. The final level of invariance is strict invariance which concerns the group equality in residual variance in addition to measurement structure, factor loadings as well as intercepts. Each level of MI is assessed via the comparison between a less constrained model that achieved the lower level of invariance and a more restrictive one. That level of invariance is established if 1) a likelihood ratio test (i.e., χ^2 test) does not suggest a model rejection, or 2) the change in model fit indexes (e.g., CFI, RMSEA) does not exceed an acceptable/pre-defined threshold, or 3) no serious change is suggested in the model modification index (e.g., Cheung & Lau, 2012).

1.2 Issues in Measurement Invariance Assessment

Despite the importance of establishing full MI, it is often found unrealistic to fulfill all levels of invariance (i.e., structural, weak, strong, strict) in practical applications (e.g., Millsap & Kwok, 2004; Millsap & Meredith, 2007). The progressive escalation in model constraints often leads to a high model rejection rate even if the group difference is small, especially when the model structure is complex (e.g., multi-factors, correlated errors) or a scale contains a large number of items. To date, the majority of MI studies have been focusing on MI testing, either by assessing the feasibility of invariance through model comparisons (Byrne et al, 1989; Cheung & Rensvold, 2002), or identifying non-invariant items (e.g., DIF testing, Kim, Yoon, & Lee, 2011; Millsap, 2011; Verhegan et al., 2016; Bayesian method, Shi et al., 2017). Comparatively, there is less attention that has been put on potential solutions when full MI cannot be achieved.

As stated in a review by Vandenberg & Lance (2000), there are two primary roles that MI assessment serves. One is in determining the quality of scales or tests, where the item that fails to pass MI testing is either excluded to maintain the homogeneity of the scale among different populations or flagged to explore the qualitative difference between groups (Cheung & Lau, 2012; Cheung & Rensvold, 1998). Another is in facilitating the subsequent analysis such as testing mean difference, selecting candidates, or making diagnosis (e.g., Lai et al., 2017; Millsap, 2011). To achieve either goal, identifying the exact non-invariant items is necessary unless a strict, or at least strong MI is attainable. However, locating the exact non-invariant item is not an easy task since the common MI assessment only is an omnibus test where the exact cause of non-invariance cannot be located. In addition, even with knowing which items are non-invariant, researchers who are wishing to utilize measurement models for predictions still need to solve for the issue of non-invariance. It is simply because that

when the non-invariance is ignored and two groups are forced to be equal, we may expect to see biased estimates on factor scores due to the model misspecification, which consequently lead to poor predictions.

1.3 Solutions to Common Issues in MI

As noted, achieving full MI is hardly ever attained in the real world. In applied settings, researchers often find themselves in situations where some but not all items are invariant for a given scale. This common situation, when full MI fails, but some items are invariant is referred to as partial invariance, which has been found as a useful alternative solution to full MI. Two components are key to partially invariant models, one is locating the exact non-invariant items, while the other is finding an appropriate model fitting strategy. Next, we will discuss some common approaches with partially invariant models, followed by discussing its pros and cons.

1.3.1 Partial MI Model

With the failure of MI, one popular choice is to identify non-invariant items and fit a partially invariant model (e.g., Cheung & Lau, 2012; Lai et al., 2017). Partially invariant models refer to the situation where some but not all the parameters (e.g., factor loadings, intercepts) hold for invariance between groups (e.g., Millsap & Kwok, 2004; Shi et al., 2019). For instance, a partially strong invariant model describes where all factor loadings and factor structure hold for invariance, while one or more item intercepts are non-invariant between groups (i.e., the number of non-invariant intercept cannot exceed the total number of items). Several model fitting strategies are considered when measurement models only hold for partial invariance. Several approaches have been discussed in the previous literatures: 1) Deleting non-invariant items; 2) Relaxing the constraints on non-invariant parameters; 3) Ignoring the non-invariance in model to fit a fully constrained model; 4) Using composites

score instead of factor models. The first option can only be achieved if there are enough items, which is not always feasible (e.g., CESD-9, PHQ-8). The second one is fitting a so-called "partially constrained model", where the group constraints are released on non-invariant parameters. This approach usually leads to the most accurate estimates given the model is correctly specified, yet sometimes is being criticized as equivalent to comparing "apples" with "pears". The third option has been shown to produce biased model parameters (e.g., factor means, regression path coefficients) in certain conditions given it is essentially fitting a statistically incorrect model (Guenole & Brown, 2014; Hsiao & Lai, 2018; Shi et al., 2019). The last option, which uses composite scores, while seemingly simple, requires even a higher level of invariance (i.e., tau equivalence). It can only be adopted when all the scale items are parallel in the population (e.g., Little, Rhemtulla, Gibson, & Schoemann, 2013), in other words, MGCFA models need to achieve an even more restrictive invariance state compared to strict invariance.

All these approaches have certain pros and cons which provide theoretical foundations for researchers when deciding which model solution to take. Yet, there are still some unsolved issues. One theoretical issue is whether a latent construct that is partially invariant still can be interpreted as representing the same construct between groups (e.g., can we compare plutos with plums, or apriums with apricots?). Another issue is that the estimation bias of different model fitting strategies can vary case by case: That is, the amount of bias in parameter estimates is subject to a set of factors such as sample size, the pattern and magnitude of non-invariance, and scale length. For instance, when the magnitude of non-invariance is small, the difference in estimation bias among different model fitting strategies (e.g., fully constrained model versus partially constrained model) may be quite trivial (Shi et al., 2019); when the magnitude of non-invariance is small, deleting the item or releasing its

equivalent constraint may not lead to a very different result compared to fitting a fully constrained model. As such, each of these model fitting strategies may be preferable to the others under certain conditions (Shi et al., 2019) (e.g., better model fit, higher more predictive accuracy, more theoretically sound). There are also some situations where different model fitting strategies are indistinguishable (e.g., produce similar estimates, fit data equally well). However, no matter which approach to partial MI is taken, the information contained in other models (those that were deemed as inappropriate) will be ignored. As such, the selection of the approach to solving the partial MI issue unwittingly introduces model selection. The "best" model is then considered as having been the true data generating process and the one producing statistical inference for the population. This brings up a long-standing issue in model selection – the scientific inference conditional on one single model can be quite limited since other possible models are overlooked (e.g., Kaplan & Lee, 2015, 2018; Madigan & Raftery, 1994; Raftery, 1995; Raftery, Madigan, & Hoeting, 1997). This consequently leads to the underestimation of the uncertainties in the parameters of interests (e.g., factor scores).

1.3.2. Locating Non-Invariant Items

As noted earlier, full MI often fails to be attained in empirical studies (Cheung & Lau, 2012; Vandenberg & Lance, 2000), and hence researchers may alternatively choose to remove non-invariant items, fit a partially constrained model, or interpret non-invariance meaningfully (Cheung & Rensvold, 1998). Either way, identifying non-invariant items is inevitable. To date, most of MI assessments are model/scale level analyses (e.g., MGCFA model comparison, likelihood-ratio test (LRT)) (Zumbo & Koh, 2005; Jöreskog, 1971), where model constraint is required, and each level of invariance is assessed as whole. The rejection of null hypothesis (i.e., invariance fails) can only indicate the existence of group

differences, but not the location of those differences (i.e., loading, intercept, residual). Therefore, to locate non-invariant items or evaluate item-level invariance, researchers need additional analyses. One approach is using the model modification index as guidance to search for the source that causes non-invariance (Byrne et al., 1989; Shi et al., 2017; Yoon & Millsap, 2007). The value of the modification index gives the expected drop in χ^2 value of a likelihood ratio test when a parameter constraint is relaxed (Muthén & Muthén, 1998-2011; Yoon & Kim, 2013), and a large value indicates that the group heterogeneity may exist in that parameter causing model misfit. Thus, to search for non-invariant items, each parameter constraint with a large modification index value is removed, one at a time, while holding other parameter constraints unchanged. This search will continue until no significant change is detected by the modification index (Cheung & Lau, 2012; Yoon & Millsap, 2007). There is a comparable approach under the item response theory (IRT) framework where multiple model comparisons are conducted via LRT. Similarly, cross-group difference/non-invariance for each parameter is not directly assessed, and the detection of non-invariant items is based on the amount of change in model fit. Several issues should be noted here: 1) When freeing one parameter constraint, the other constrained parameters must be assumed truly invariant; 2) These approaches result in the multiple comparison problem, leading to inflated type I and II errors; 3) As the number of scale items increases, the complexity of the model structure may also increase (e.g., correlated error variances, cross-loading items), which not only amplify the issues in 1) and 2) but also make the searching procedure more cumbersome.

Alternatively, researchers could choose item-level MI assessments where assessing invariance and locating non-invariance are performed simultaneously (e.g., Cheung & Lau, 2012; Kim, Yoon, & Lee, 2011; Millsap, 2011; Verhagen, Levy, Millsap, & Fox, 2016; Zumbo & Koh, 2005). Item-level MI assessments focus on the group difference/non-

invariance in item parameter values (Millsap, 2011) and assess the group difference in each parameter value directly. Full MI is established if group differences are not detected in any of the item parameter values. If any of the parameter values is found different on a single item, this item will be considered to be non-invariant thus leading to the rejection of full MI. Several methods have been developed to assess item-level MI under both MGCFA and IRT frameworks (Cheung & Lau, 2102; Millsap, 2011; Shi, Song, DiStefano, Maydeu-Olivares, McDaniel, & Jiang, 2019; Thissen, 1982; Verhagen et al., 2016). In MGCFA, Cheung & Lau (2012) proposed to directly assess the item parameter difference through bootstrapping confidence interval using maximum likelihood estimation. In IRT, a Wald test is widely applied (Thissen, 1982) for assessing item-level invariance. Alternatively, Shi et al., (2019) used a Bayesian SEM to assess each parameter difference via the credible intervals (CrI's) of the posterior distribution. Instead of using CrI's, Verhagen et al., (2016) developed a Bayesian method to assess item parameter differences using Bayes factors (BF) (Kass & Raftery, 1995).

Among all the aforementioned methods, the most straightforward approach of detecting non-invariant items is to directly compare the parameter estimates of item loadings, intercepts, and residual variances between groups (Millsap, 2011, p.79). Non-invariance is detected if any of these comparisons pass a threshold of a given statistic (e.g., $p < .05$, BF > 3). In the Bayesian framework, these between-group comparisons can be simplified by estimating a set of hyperparameters that are defined to represent the cross-group differences (Pokropek, Schmidt & Davidov, 2020). In other words, the core research question comes down to identifying which of these hyperparameters are statistically meaningful. From this viewpoint, the detection of non-invariant items becomes a variable selection problem, which can be solved by utilizing some popular variable selection methods. To date, the connection

between locating the exact non-invariant items and variable selection has not been explicitly mentioned or investigated. This connection will be discussed here in the hopes of expanding possibilities for dealing with partial measurement invariance.

1.4 Model Selection and Model Averaging

Selecting the best model fitting strategy with partially invariant models has proven to be difficult in both computational and theoretical perspectives. Another aspect to keep in mind is that no matter which model fitting strategy is being selected, we are making inferences about population conditional on one single model. The fundamental issue of making decisions from a single model is that it ignores the uncertainty in model selection and holds the belief that the final selected model represents the true data generating process (e.g., Kaplan & Lee, 2018; Madigan & Raftery, 2012; Raftery, Madigan & Hoeting, 1997). This is in line with the now famous quote by Box (1979), who noted that "All models are wrong, but some are useful."

In actuality, we never know the true data generating process and pretending a single model solution over many competing models is the true one can result in too much certainty (e.g., Navarro, 2019). This risk of being overly confident in the inference and decisions made from the "best" model solution is often underestimated (Hoeting, Madigan, Raftery & Volinsky, 1999). This is especially the case in social sciences where the scope of most studies is on the population level. For example, in SEM applications, the methods for model comparison and selections are well studied (Lin, Huang, & Weng, 2017; Liang & Luo, 2019), and researchers are also guided by substantive theories when choosing a single model fitting strategy. Yet, whichever model is selected or decided to be the best is conditional on current data set and the past theories. Consequently, the complexity of human characteristics and the evolvement of scientific theories are largely overlooked. In other words, whether a single

model solution conditional on a given data can generate a better prediction for unseen data or can be applied on the future empirical studies is under doubt (James, Witten, Hastie, & Tibshirani, 2013). In partially invariant models, the pattern of non-invariance can be complex and often varies case by case (Lai et al., 2017), which impose many uncertainties during model selection process. Usually, selecting one model solution (e.g., omitting non-invariant items) leads to the disacknowledgment of other possibilities which may contain important information about the population. For instance, non-invariant parameters could be manifest as invariant under certain conditions but not others (e.g., small N versus large N, Meade & Bauer, 2007). For selection, comparison, or diagnosis, researchers would require factor scores computed from a partially invariant model that are not only suitable for the current dataset, but also applicable to future studies. Accounting for model uncertainty and combining all possible models may be a better solution for partially invariant models (Kaplan & Lee, 2018; Madigan & Raftery, 2012; Raftery et al., 1997). Thus, the current dissertation proposes to take a model averaging approach as an alternative solution to selecting one single model fitting strategy.

One way of incorporating non-invariance in the measurement model is including all possible model sets in the estimation process and get the averaged results via Bayesian method, namely, Bayesian Model Averaging (BMA) (e.g., Kaplan & Lee, 2015, 2018; Raftery et al., 1997). Bayesian statistics is known for handling the uncertainty in parameter estimation, and BMA solves one more layer of uncertainty, the uncertainty during the model selective process (Madigan & Raftery, 1994; Raftery et al., 1995, 1997). Rather than settling on one single model, BMA considers all models that are deemed to be theoretically and scientifically possible, and hence takes care of the uncertainties in both parameter estimates as well as modelling process by computing different model probabilities before averaging

across all models. In comparison to a single model solution, BMA preserves information from all possible model sets while giving those models different weights, and hence possesses the ability of optimizing out-of-sample predictions. This is in line with researchers who argue that our end goal of MI assessment is not simply evaluating invariance but utilizing the measurement model for practical purposes (e.g., group comparison, candidate selections) (Hsiao & Lai, 2018; Lai, Kwok, Yoon, & Hsiao, 2017; Millsap & Kwok, 2004). In both simulation and empirical studies, BMA has shown its excellent predictive ability in various modelling techniques (e.g., graphical models, SEM, linear regression; Kaplan & Lee, 2015, 2018; Madigan & Raftery, 1994; Raftery et al., 1995, 1997).

Nevertheless, implementing BMA has been found challenging because it requires averaging over the entire model space and the number of possible candidate models in the model space often varies across different modeling situation (e.g., Kaplan & Lee, 2015, 2018; Raftery et al., 1997). In regressions, this means including all the possible combinations with each predictor. With partially invariant models, this means including all the possible combinations of different measurement models with each of the non-invariant parameter maybe constrained, freely estimated, or fixed to zero/omitted (e.g., partially constrained model, fully constrained model). Additionally, the amount of non-invariance can vary across non-invariant parameters, which means that they could be negligible for some parameters, while highly influential for others. Hence, possible candidate models could also be each/some non-invariant parameter constrained at a time while other ones freely estimated. Consequently, the number of possible modelling strategies for a partially invariant model can grow exponentially. Therefore, using BMA with accounting for all possible situations is unfeasible in practice because the model space can become enormous. To overcome this computational difficulty of BMA and reduce the size of candidate model sets, researchers

have proposed some model selecting rules such as "Occam's window" to exclude some unnecessary models (Kaplan & Lee, 2015, 2018; Raftery et al., 1997). For the MI question, using "Occam's window" will first require the identification of non-invariant parameters, and then implementing the BMA to get averaged estimates. Alternatively, there is a simpler solution: utilizing appropriate priors that are known to select specific variables via regularization of the whole parameter space (such as the horseshoe prior) to identify non-invariant parameters and get averaged estimates from all possible models simultaneously.

1.4.1 Incorporation of Non-invariance in the Measurement Model using the HS Prior

Here, we will focus on a specific prior, the so-called horseshoe (HS) prior (e.g., Carvalho, Polson, & Scott, 2010). The HS prior originates from a multivariate-normal scale mixture distribution and has a "horseshoe" like shape where its left side distribution with no shrinkage handles signals while the right-side distribution with almost a complete shrinkage handles noise (Carlos, Carvalho & Nicholas, 2010). This nice statistical property of HS prior allows it to function as a parameter "switch" and produces BMA like estimates (Carvalho, Polson, & Scott, 2009). As mentioned, using the BMA in estimating partially invariant models requires getting a candidate model set which contains all the possible modelling strategies. The candidate model set for a partially invariant model is a special case because all the possible models are nested within each other. In other words, there will be no new variable introduced in the candidate model set and all the candidate models could be easily transformed from one to another by adding or omitting some parameters. Specifically, all the models within a candidate model set can be technically transformed from each other by fixing relevant parameters to certain values (e.g., zero). For instance, we can turn a partially constrained model into a fully constrained model by fixing non-invariant parameters to be the same across groups. We can also obtain a shortened measurement model with all the non-

invariant scale items deleted by fixing their parameters (i.e., loading, intercept, residual variance) to be zero. This model transformation process can be cumbersome under the frequentist statistic framework, since we need to fit a sequence of models where the number of models will rapidly increase as the complexity of models and non-invariant situations increase. In the Bayesian framework, these model transformations can be easily handled by placing the HS prior on the hyperparameters that are defined as cross-group differences. Along with the target dataset, the model shifting direction (i.e., non-invariant versus invariant) and quantity are automatically driven by researchers' prior beliefs (i.e., some but not all the parameters are completely invariant) as well as the empirical evidence (i.e., data). Specifically, the right side of HS prior handles the parameters with negligible differences, while the left side HS prior takes care of the parameters with notable differences. Hence, the same result of averaging over the entire model space is obtained by using the HS prior to turn equality constraints off and on based on data. In estimating partially invariant models, the HS prior becomes a natural substitute for BMA.

1.4.2 Locating Non-invariant Items using the HS Prior

One way of locating the exact non-invariant items is to directly assess the cross-group difference on item parameters (i.e., loading, intercept, residual variance). In Bayesian statistics, we can simply define a set of hyperparameters to represent the cross-group difference on each item parameter. The task of locating non-invariant items is to assess if any of these "differences" parameters are effective or not. For the most established measurements, it is reasonable to assume that the number of invariant items exceeds those non-invariant ones. To locate non-invariant items, we need to search for sparse signals among numerous noises. Therefore, we can reframe the locating of non-invariance as a signal detection problem, where the invariance is obtained in the absence of signal, while the non-

invariance is detected in the presence of signal. In the statistical machine-learning literature, methods in handling sparse signals and identifying relevant variables have been well developed (Carvalho, Polson, & Scott, 2009, 2010; Piironen, & Vehtari, 2017), and some of them can be useful in locating non-invariances. The sparse estimation issue arises when a statistical model contains a large number of parameters, yet only few of them are expected to be relevant, or in other words, have coefficients that significantly depart from zero (e.g., Piironen, & Vehtari, 2017a, b). The task under this situation is to detect signals in the presence of noise. Several solutions are proposed under both Frequentist and Bayesian frameworks. In the Frequentist framework, Lasso regression and some of its generalized versions are used where the sparse signals are handled by penalty parameters (James, Witten, Hastie, & Tibshirani, 2013; Piironen, & Vehtari, 2017; Tibshirani, 1996). In the Bayesian framework, two families of priors: two components discrete mixtures prior and shrinkage priors, are commonly used to handle sparsity (Carvalho, Polson, & Scott, 2009; Piironen, & Vehtari, 2017). Given that defining non-invariance as a hyperparameter can only be achieved using Bayesian estimation method, we will not go into details about variable selection methods under a Frequentist framework. Two-component discrete mixture priors, which are also known as Spike-and-Slab priors, use a point mass centered at zero to describe irrelevant variables, and an absolute continuous space to capture variables that are deemed to be non-zero (Ishwaran & Rao, 2005; Mitchell, & Beauchamp, 1988). Shrinkage priors on the other hand, tend to handle the sparsity by compressing the total effect of all predictors with a continuous "shrinking" distribution centered close to zero (e.g., Tibshirani, 1996). From the theoretical perspective, the Spike-and-Slab prior should be a more appropriate choice, given it perfectly mimics sparsity by using the "spike" to represent the absence of signals, and the "slab" to represent the presence of signals (Carvalho, Polson, & Scott, 2009; Mitchell & Beauchamp, 1988). Yet, several estimation issues arise when implementing two component

discrete mixture priors in complex models (e.g., containing a large number of parameters to be estimated) which makes it difficult to operate in practice (Piironen & Vehtari, 2016). Alternatively, as a member of shrinkage priors, the HS prior highly resembles the Spike-and-Slab prior in the distribution shape but does not contain a point mass at zero which may result in estimation problems when using methods such as Hamiltonian Monte Carlo that require smooth and differentiable posterior distributions. As such, the HS prior can be seen as a more generally applicable substitute to the Spike-and-Slab prior and, for this current work, a very promising approach for locating non-invariant items.

1.5 Overview of Dissertation

Numerous research papers have highlighted the importance of achieving MI for social and behavior research. Nevertheless, comparing to the large body of MI studies focusing on the assessment of MI, there are fewer studies that are concerned with the failure of invariance and the elaboration of alternative solutions. As noted, the probably most popular choice is to fit partially invariant models. But this approach requires the exact identification of non-invariant parameters and leaves the question open about which model fitting strategies should be selected. Thus, the inspiring questions behind this dissertation are; (a) How do we locate the exact item-level non-invariance/invariance? (b) Can we incorporate the non-invariance into the partially invariant measurement model using sparse Bayesian estimation methods (i.e., HS prior) to improve the model prediction; and (c) How do we apply these techniques using empirical data? Thus, the goal of current dissertation is threefold: First, we discuss and demonstrate how to use the HS prior for detecting non-invariant/invariant items via simulated data example under different conditions. Second, we evaluate the predictive performance of using the HS prior with partially invariant measurement models and compare its performance with other alternative methods. Third, we use empirical data to demonstrate how to use the

HS prior to identify the exact non-invariance and get the factor scores of a partially invariant model for a candidate selection.

2. Chapter 2

Detecting Non-invariant Items

2.1 Current approaches in detecting non-invariant items

Generally speaking, there are two classes of statistical approaches for detecting non-invariant items: (1) model comparison approaches, and (2) item-level analysis approaches. Model comparison approaches are commonly applied in multigroup confirmatory factor analysis (MGCFA) under the frequentist statistical framework. This approach assesses MI on a model-level which does not examine the state of invariance for each item separately, and thus usually requires a sequence of constrained model comparisons for identifying all non-invariant items. Item-level analysis approaches have been applied in MGCFA as well as item response theory (IRT), both under frequentist statistical frameworks and Bayesian statistical frameworks. In contrast to model comparison approaches, item-level analysis approaches directly focus on the cross-group difference of each single parameter, so that all the non-invariant items can be identified at the same time. In the following two sections we will review these methods in frequentist and Bayesian statistics.

2.1.1 Frequentist approaches in detecting non-invariant items

The most common way of identifying non-invariant items under the frequentist statistical framework is through MGCFA (e.g., Millsap, 2011; Schmitt & Kuljanin, 2008; Vandenberg & Lance, 2008; Jung & Yoon, 2016, Zumbo & Koh, 2005). A standard formulation of MGCFA can be written as:

$$X_k = \mu_k + \Lambda_k \Psi_k + \epsilon_k \quad (1.1)$$

where X_k is an $p \times 1$ vector of the observed item scores, μ_k refers to a $p \times 1$ vector of the item intercepts, Λ_k denotes a $p \times m$ matrix of the item factor loadings, W_k represents an $m \times 1$ vector of the latent factor scores ($p > m$), and ϵ_k stands for a $p \times 1$ vector of the unique factor scores. The group membership is denoted by the subscription k which indicates that each parameter in equation (1.1) can vary across groups. The assumption of independence between the latent factors and unique factor scores (i.e., $cov\ (\Psi_k, \epsilon_k) = 0$) leads to the variance-covariance of X_k

$$\Sigma_k = \Lambda_k \Phi_k \Lambda'_k + \Theta_k, \tag{1.2}$$

where Σ_k denotes a $p \times p$ variance-covariance matrix of X_k, Φ_k stands for a $m \times m$ variance-covariance matrix of Ψ_k and Θ_k refers to a $p \times p$ variance-covariance matrix of ϵ_k. It should be noted that Θ_k is usually a diagonal matrix containing the variances of ϵ_k assuming independence among all unique factors (i.e., $cov\ (\epsilon_k, \epsilon_k) = 0$). The mean structure of X_k can be expressed as

$$E(X_k) = \Lambda_k \Psi_k + \mu_k \tag{1.3}$$

where Ψ_k is a vector of factor means for the latent variable W_k, so that the likelihood of the observed values is $X_k \sim N(E(X_k), \Sigma_k)$.

Assuming MI assessment is considered between two groups (i.e., focal group versus reference group), each level of invariance testing can be proceeded in the following sequence: configural invariance (i.e., same pattern of the measurement model), weak

invariance (i.e., $\Lambda_f = \Lambda_r$), strong invariance (i.e., $\Lambda_f = \Lambda_r$, $\mu_f = \mu_r$), and strict invariance (i.e., $\Lambda_f = \Lambda_r$, $\mu_f = \mu_r$, $\Theta_f = \Theta_r$). The variance-covariance matrix and mean structure of a strictly invariant model can be written as:

$$\Sigma_k = \Lambda \Phi_k \Lambda' + \Theta \tag{1.4}$$

$$E(X_k) = \Lambda K_k + \mu. \tag{1.5}$$

Equation (1.5) implies that the observed group mean difference is mostly driven by the difference in factor scores. Each level of invariance is tested through a series of model comparisons where a more constrained model (e.g., $\Lambda_f = \Lambda_r$,$\mu_f = \mu_r$) is compared with a less constrained one (e.g., $\Lambda_f = \Lambda_r$). If these two models significantly differ in terms of goodness-of-fit (e.g., Δ CFI > .01, Δ RMSEA >.01, $P(\chi^2)$ <.05) (e.g., Byren et al., 1989; Cheung et al., 2002; Millsap, 2011), then one can conclude that MI fails for the more constrained level, and only holds for the less constrained one. MI can fail at any stage during an invariance testing due to one or more non-invariant items. Yet, searching for non-invariant items happens only after this sequence of model comparisons.

Under MGCFA, three statistical methods are commonly applied when searching for non-invariant items: 1) the model comparison approach via the likelihood ratio test (i.e., χ^2 test), or using the change in comparative fit indices (e.g., Δ CFI) (Rensvold & Cheung, 1998), 2) the sequential search approach using model modification indices (Yoon & Millsap, 2007) , 3) the parameter comparison approach using the bias-corrected bootstrap confidence interval (CI) method (Cheung & Lau, 2011; Jung & Yong, 2016, Meade & Bauer, 2007). In the model comparison approach, the same logic as MI testing is applied. That is, when a MI assessment fails at a certain stage of invariance testing (i.e., weak, strong, strict), a sequence of model comparisons will be performed between a MGCFA model that is constrained at that level (i.e., $\Lambda_f = \Lambda_r$ or, $\mu_f = \mu_r$, or, $\Theta_f = \Theta_r$) and a set of identically constrained models

which have one parameter relaxed from equality constraint at a time. If any of these less constrained models does not fit significantly worse compared to the more constrained model (e.g., Δ CFI > .01) (Cheung & Lau., 2011, Cheung & Rensvold, 2002), then the item associated with that freely estimated parameter is flagged as non-invariant. For instance, to locate non-invariant items for a MGCFA model that fails on the strong invariance testing (i.e., $\Lambda_f = \Lambda_r$, $\mu_f = \mu_r$) a sequence of model comparisons will be performed between a MGCFA model that is constrained on both factor loadings as well as intercepts and a set of models with each item intercept freely estimated at a time. If any of these models which has a freely estimated intercept fits significantly better than the more constrained model, then the item association with that intercept is deemed to be non-invariant. Despite its wide application, several limitations of this model comparison approach should be noted. One major drawback of this approach is the use of various evaluation criteria such as $\Delta\chi^2$, and ΔCFI. Similar to the χ^2 index that is used to assess the entire model fit, $\Delta\chi^2$ is also sensitive to sample size. In other words, a significant $\Delta\chi^2$ may not only be caused by a notable difference between two models but could also be a result of a large sample size. In terms of ΔCFI, which has no known sampling distribution and thus is not subjectable to any significance testing, the use of a predetermined cutoff value (e.g., .01) is often being criticized as arbitrary (Cheung & Lau, 2012) and unstable when evaluating models with different levels of complexities (Cheung & Rensvold, 2002). Additionally, multiple model comparisons not only make the searching procedure cumbersome (especially for long scales), but also lead to an inflated type II error rate given that the invariance/nonvariance of each parameter is conditional on the "true" invariance status of other parameters.

In the sequential search approach, the model modification index of a constrained baseline model is us to identify non-invariant items. Specifically, for a MGCFA model that

fails at a certain stage of invariance testing, invariance constraints of each parameter will be sequentially removed based on their modification indices to achieve a better model fit, starting from the one with the largest value until no index exceeds a pre-determined cutoff value (e.g., 5) (Yoon & Millsap, 2007; Jung & Yoon, 2016). Therefore, parameters are deemed to be non-invariant if their model modification indices pass certain pre-determined values. This method seems quite promising and easy to operate given that model modification indices are readily available in most commercial/non-commercial software such as Mplus (Muthén& Muthén, 1998-2011) and lavaan (Rosseel, 2012), yet it is built upon a strong assumption that all the constrained parameters are truly invariant. Only when this assumption holds, the values of model modification indices are reliable. Ironically, this is not a testable assumption.

The parameter comparison approach is an item-level analysis (Cheung & Lau, 2012; Meade & Bauer, 2007; Zumbo & Koh, 2005), which does not require imposing invariance constraints on model parameters, and thus is able to examine the invariance of all items simultaneously. Precisely, this approach proposed using 1000 bootstrap samples created by modern statistical software (e.g., Mplus) to compute a set of parameters (i.e., $\Delta\lambda_{r-f}$, $\Delta\tau_{r-f}$) representing cross-group differences in factor loadings and intercepts (Cheung & Lau, 2012; Meade & Bauer, 2007). An item is deemed to be noninvariant if the bias-corrected CI's of any of their Δ parameters do not include zero. Compared to the previous two methods, the parameter comparison approach simplifies the multiple testing steps to one and eliminates the need for estimating constrained models that often lead to non-convergence issues. However, because multiple hypotheses are tested at the same time, this method does require an adjustment for the nominal α level to minimize the overall Type I error rate.

2.1.2 Bayesian approach in detecting non-invariant items

Under the frequentist framework, the goal of MI testing is to achieve exact invariance and the decisions of MI testing are usually dichotomous. Thus, there are no cross-group differences allowed when estimating MGCFA models. Some recent studies under the Bayesian framework argued that an absolute invariance can hardly be achieved in practice and allowing a small "wiggle room" on group constraints is a more reasonable option (Muthén & Asparouhov, 2012, 2013, 2017). Therefore, the concept of approximate invariance was proposed where small variance priors (e.g., an area ~ N (0,.05)) are utilized to represent approximate equality constraints under the Bayesian SEM (BSEM), (Muthén & Asparouhov, 2012, 2013, 2017; Pokropek et al., 2020; Van de Schoot, Kluytmans, Tummers, Lugtig, Hox, & Muthén, 2013). Along with this Bayesian approach, researchers have developed an item-level analysis under MGCFA to identify non-invariant items via evaluating posterior distributions of parameter differences (e.g., $\Delta\lambda_{r-f}$) (Shi, Song, DiStefano, Maydeu-Olivares, McDaniel, & Jiang, 2019, Shi, Song, Liao, Terry, & Snyder, 2017). Like in the parameter comparison approach, a set of Δ parameters are pre-defined and serve as a threshold that allows certain parameter differences. The models are then estimated under approximate invariance constraints (i.e., small variance priors) in Bayesian MGCFA (Shi et al., 2017, 2019). Two assessment criteria are used in deciding non-invariant items: 1) 95% credible interval (CrI), 2) 95% highest density interval (HDI). The rule of thumb states that if the 95% CrI of posterior distributions of any parameter difference excludes zero, or the 95% HDI of posterior distributions of any parameter difference falls within a region that is deemed to be practically equivalent (i.e., region of practical equivalence: ROPE), then non-invariance is detected for that specific parameter (Kruschke, Aguinis, & Joo, 2012; Shi et al., 2019).

Another major critique on frequentist methods in MI assessment pertains to the null hypothesis significance testing (NHST) (Kruschke 2011; Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010). Researchers argued that NHST used by frequentist methods hardly ever provides evidence in favor of invariance given that it infers non-invariance through the rejection of null hypotheses (Verhagen, Levy, Millsap, & Fox, 2016). To overcome this issue, a bayes factor (BF) approach was proposed to gather evidence in support of invariance and non-invariance hypotheses, for each parameter in IRT modelling (Verhagen & Fox, 2013; Verhagen, Levy, Millsap, & Fox, 2016). In Verhagen et al., (2013; 2016) studies, a set of cross-group differences for each parameter of interest are pre-defined as Δ parameters, which are then being used to evaluate for invariance. Specifically, the probabilities of Δ parameters as zero are approximated using their posterior distributions ($P(\Delta_P = 0|H_1, D)$, and the probabilities of Δ parameters as non-zero are assumed under some normal condition (i.e., Cauchy distribution). Then, a Savage-Dickey density ratio is used to calculate the $BF_{01} = \frac{P(\Delta_P=0|H_1,D)}{P(\Delta_P=0|H_1)}$ (Dickey, 1971; Kass & Raftery, 1995), where the BF_{01} reflects the ratio between the probability of invariance and non-invariance for each Δ parameter given the data. Using a rule of thumb, the invariance is said to be held once a BF_{01} exceeds 3, while non-invariance is detected if a BF_{01} is less than 1/3. Notably, Verhagen and colleagues' work was the first to demonstrate the use of BFs to infer MI and to underscore its ability to obtain fine grained comparisons for a large number of relevant parameters. Verhagen and colleagues' item-level invariance testing appears to be desirable since the identifications of non-invariant items and model invariance are accomplished in one step.

One question that needs to be asked is what should be the best choice of prior distribution on Δ parameters in the Bayesian MI? Whether the choice of the prior distributions on Δ parameters will impact the detection of non-invariant items is

inconclusive (Shi et al., 2017, 2019; Verhagen et al., 2012, 2016). Shi et al., (2017, 2019) have examined the influence of using highly informative priors (i.e., N(0, 0.001), N(0, 0.01), N(0, 0.05), N(0, 0.1)) and non-informative prior on detecting non-invariant items and found that the difference is minor. In Verhagen and colleagues' studies (2012, 2016), two different prior distributions (i.e., standard normal distribution, Cauchy distribution) have been used to obtain BFs for detecting the non-invariant items. They found that under a small sample size condition, a more informative prior could lead to a higher Type II error rate, which means that non-invariant items are more likely to be falsely identified as invariant (i.e., false negative). The choice of priors should reflect researchers' belief about the reality to some degrees. In the context of detecting non-invariant items, the goal is to determine the invariance status of items: invariant or non-invariant. However, the priors have been used for Δ parameters in previous studies such as N(0, 0.01), or N(0, 1), is either too close to invariance, or way too diffuse to stand for any belief, which may be computationally advantageous, but not theoretically sound. Here, we believe that a two components mixture prior (Spike-and-Slab) or its closely related prior such as horseshoe (HS) could be a better choice in detecting non-invariant items.

This section has summarized some existing methods that have been developed for locating non-invariant items in most psychological research. The next section will introduce a Bayesian method for detecting non-invariant items from a novel perspective – Bayesian variable selection.

2.2 Proposed Methodology: Using Bayesian Variable Selection Methods to detect non-invariant item

The previous section has shown that the most straightforward way of locating non-invariant items is directly assessing the parameter difference of each item. In a frequentist

framework, the parameter difference is usually calculated after estimating MGCFA models, which takes two steps. In contrast, within a Bayesian framework, the parameter difference can be defined beforehand and estimated in one step. Additionally, there is no distribution assumption needed for significance testing within Bayesian framework. Yet, one question that remains is how to select an appropriate prior for the difference parameter. Priors have been used in previous Bayesian MI research mainly focused on the computational part of the model estimation. The conceptual part of prior selection for identifying non-invariant items has rarely been discussed. For well-established scales or measurements, it is reasonable to assume that most scale items operate in the same way for participants with different group memberships or being measured at different time points. In other words, the number of non-invariant items is expected to be much smaller compared to the number of invariant items. Thus, the detection of non-invariant items problem can be viewed as a classical variable selection problem, where there is a large number of predictors yet only some of them are deemed to be relevant (e.g., Ishwaran & Rao, 2005). This requires finding a sparse solution where only some of the difference parameters are effective (Carvalho, Polson & Scott, 2009). Therefore, we need to find specific kinds of priors that are able to pick up signals (non-invariance) while allowing noise (invariance). These kinds of priors are often being used in solving Bayesian variable selection problem/ Bayesian sparse learning cases where only some of the predictors in a regression model are considered to be relevant. The logic of variable selection can be applied to the identification of non-invariant items. In this dissertation we will propose a Bayesian variable selection method for identifying non-invariant items. In the remainder of this chapter we will discuss: First, how to approach the identification of non-invariant items as a variable selection problem; next we will describe how to identify non-invariant items within a Bayesian framework using a horseshoe (HS) prior used in Bayesian model selection; Finally, we will conduct a simulation study to

investigate the performance of HS priors in identifying non-invariant items under various conditions.

2.2.1 Reframing Non-invariance as the presence of an effect

To test if cross-group differences are present, we can introduce a variable or a set of variables to stand for the cross-group difference of each item in factor loadings, intercepts as well as residual variances, and conduct statistical testing to see if any of their coefficients deviate from zero. In the context of identifying non-invariant items, the absence of an effect would indicate invariance, while any significant deviations from zero of an effect would otherwise suggest non-invariance. Therefore, we can achieve the goal of detecting non-invariant items as identifying effective predictors in a linear regression fashion.

2.2.2 Using hyperparameters to define cross-group differences

In a frequentist statistical framework, defining variables that capture cross-group differences is often done by taking the parameter difference after the MGCFA model has been estimated (Cheung & Lau, 2011; Jung & Yong, 2016, Meade & Bauer, 2007). Yet, it is not easy to conduct statistical testing for these "difference" variables given that they don't have standard errors and thus some additional statistical adjustments are also needed (e.g., generate bootstrapped confidence intervals). Within a Bayesian statistical framework, this issue can be directly addressed by using hyperparameters to represent cross-group differences and the estimation process can be reduced to one step. We hereby use a set of Δ parameters to represent cross-group differences of factor loading (Δ_λ), item intercepts (Δ_τ), and residual variances (Δ_ϵ). The remaining question is how to select an appropriate prior distribution for these Δs.

2.2.3 Prior choice on the cross-group difference

Now, return to the question of detecting non-invariant items, the hypothesis being tested is whether a cross-group difference is present or not. That is, the prior distribution of the Δ coefficients should be chosen in a way to support this dichotomized decision. While we expect that most of the Δ‘s should be close to zero, indicating invariance, the prior distribution needs to allow some signals large enough to be detectable in order to detect non-invariance. A closely related question has been studied in the context of Bayesian sparse learning, where the goal is to identify a small number of relevant predictors from lots of irrelevant ones (Carvalho, Polson & Scott, 2009; Ishwaran & Rao, 2005; Piironen & Vehtari, 2017). From a Bayesian learning point of view, there are two families of prior distributions that have been discussed in solving sparse estimation problems: shrinkage priors and two component discrete mixtures (Carvalho, Polson & Scott, 2009). Shrinkage priors model the effects of predictors using a complete continuous distribution that approximates zero. While two component discrete priors, which are also called “Spike-and-Slab” priors, use a point mass at zero to shrink irrelevant predictors to zero and a complete continuous distribution to capture non-zero effects. For this purpose, “Spike-and-Slab” priors perfectly mimic MI situations where most of the cross-group differences are expected to be zero and only a small number of them significantly deviate from zero and thus should be the most ideal choice. However, “Spike-and-Slab” priors are not ready for use in some popular Bayesian statistical modelling tool due to the sampling issue (e.g., Stan). Alternatively, shrinkage priors are relatively easy to implement, computationally convenient and even result in a similar performance (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2016, 2017). One notable example is the so-called “horseshoe” (HS) prior which exhibits comparable performance to “Spike-and-Slab” priors in solving sparse estimation issues. The HS prior contains some desirable properties that enable it to handle different sparse patterns while allow large effects

to manifest (Carvalho et al., 2009, 2010). Therefore, we adopt the HS prior in this work rather than the "Spike-and-Slab" prior. Next, we will describe the HS prior in greater detail and we will show how it can be used to detect non-invariant parameters within a MGCFA model.

2.2.4 Detecting non-invariance using the Horseshoe shrinkage prior

Let us start introducing the HS prior by considering a generically defined parameter difference Δ given a MGCFA model, where $(\Delta|\beta) \sim \mathrm{N}(\beta, \sigma^2\mathrm{I})$. Δ represents the cross-group differences in factor loadings, intercepts and residual variances, and β stands for the coefficient for all Δ parameters. Assuming there is a total of *K* items for three parameters (factor loadings, intercepts, and residual variances) in a scale and only few of them are expected to exhibit non-invariance, we can then set the HS prior for $\beta = (\beta_1, \beta_2, \beta_3 \ldots, \beta_{\{3*k\}})$

$$\beta_i|\,\lambda_i, \tau \sim N(0, \lambda_i^2\tau^2)$$

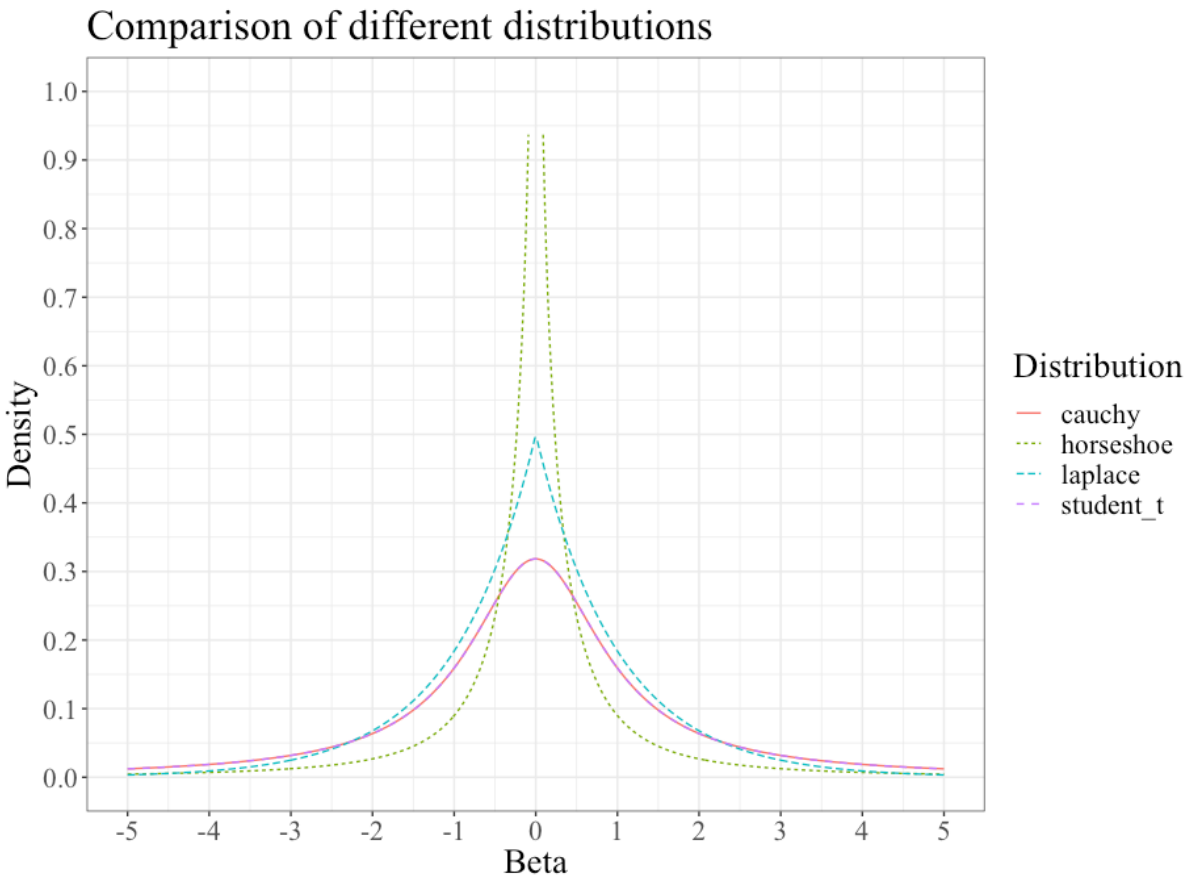
$$\lambda_i \sim C^+(0,1), i = 1, \ldots, 3 * K. \qquad (2.1)$$

Where there are total 3*K parameters to be estimated, and only few of them are deemed to be relevant. λ_i is referred as local shrinkage parameters for each β_i, τ is referred as a global shrinkage parameter as a scaling factor, and $C^+(0,1)$ is a half-Cauchy distribution describing λ_i. When handling sparsity issues, the global shrinkage parameter τ regulates all Δ towards zero, while the local shrinkage parameter λ_i permits some effective Δ_i to escape from the shrinkage. In the context of detecting non-invariant items, the HS prior's tall spike around the origin shrinks the coefficients of invariant parameters towards an infinitely small value, while

its flat Cauchy-like tail allows the coefficients of non-invariant parameters remain large (see *Figure 2. 1*) (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2016, 2017).

Figure 2. 1

The Horseshoe Prior



This framework meets the model assumption of MI where there is a small number of items that are expected to be non-invariant. In other words, some β values should deviate significantly from zero and take on meaningful quantities, while other β values, which are around zero, should be compressed towards zero. To illustrate, we borrow the idea from Carvalho et al., (2010), first assume that $\tau^2 = 1$ and define a random shrinkage weight $w_i = 1/(1+\lambda_i^2)$, which stands for the amount of weights that the posterior mean of β_i takes on zero given the data y:

$$E(\beta_i|y_i,\lambda_i^2) = \left(\frac{\lambda^2}{1+\lambda^2}\right) * y_i + \left(\frac{1}{1+\lambda^2}\right) * 0 = (1-w_i) * y_i. \quad (2.2)$$

Given w_i is bounded between 0 and 1, and thus according to Fubini's theorem,

$$E(\beta_i|y) = \int_1^0 (1-w_i) y_i \pi(w_i|y) dk_i = \{1 - E(w_i|y_i)\}\, y. \quad (2.3)$$

By examining the prior choices on w_i that are implied by different $\pi(\lambda_i)$, along with this transformation, we can then get a clear idea about the advantage of applying the HS prior to distinguish between non-invariance (signals) and invariance (noises). As in equation (2.1), $\lambda_i \sim C^+(0,1)$ suggests that $w_i \sim Be(1/2,1/2)$, a symmetric density that is bounded between 0 and 1. This horseshoe shaped distribution indicates that two things are expected in the data: Strong signals indicate non-invariances ($k \approx 0$, no shrikage), and noises indicate invariances ($k \approx 1$, complete shrinkage).

2.2.5 Using a Bayes Factor in quantifying non-invariance

To evaluate the invariance status of each parameter via the posterior distribution of βs, we need to select an appropriate decision-making tool. There are several commonly used metrics such as 95% credible intervals (CrI), Bayes factors (BF), and Bayesian predictive probabilities. Out of those, we consider a Bayes factor (BF) approach, as it has shown good performance in the evaluation of different degrees of invariance and non-invariance in a Bayesian IRT framework (Verhagen et al., 2016). Indeed, BFs have a long history of serving as a decision-making tool for model/variable selection and they convey information on how much more likely one model/hypothesis is, in relation to another one given the data.

A Bayes factor represents the probability of the data under one hypothesized mode(H_0) relative to another (H_1). Each model consists of a likelihood function and prior

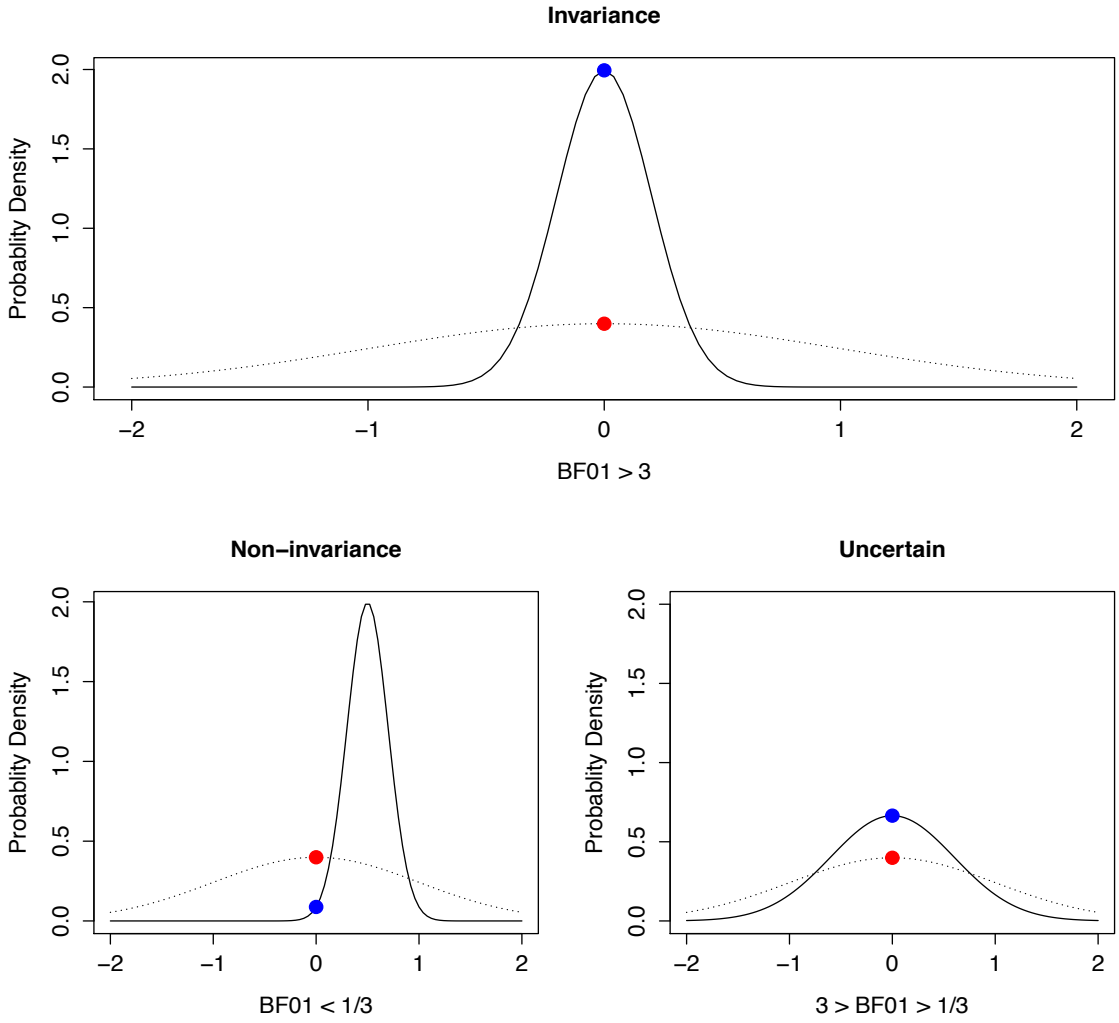
distributions over the unknown parameters. In hypothesis testing, the prior on a parameter encodes the uncertainty in the parameter given the hypothesis. For example, if H_0 posits that a parameter is exactly zero, then the prior is a point mass at zero. Alternatively, if H_1 posits that a parameter has a .95 probability of being between -1.96 and 1.96, and a .50 probability of being either positive or negative, then the prior may be based on a standard normal distribution. The likelihood of the data is integrated over the priors of each hypothesis to yield marginal likelihoods: $p(D|H_k) = \int p(D|\theta, H_k)p(\theta|H_k)d\theta$. The BF then is simply the ratio of these marginal likelihoods: $BF_{01} = \frac{P(D|H_0)}{P(D|H_1)}$. The ratio reflects the relative support of one hypothesis over another, accounting for uncertainty in the hypothesized parameters. Hence, the BF provides the evidence in form of a likelihood ratio that takes into account the data and the prior reflecting different hypothetical predictions (Wagenmakers, 2007). While the BF yields a continuous measure of the ratio, Raftery (1995) has proposed that any values of BFs larger than 3 can be deemed as supportive evidence – it's important to note that this threshold is arbitrary. The ratio of a BF reflects how many times more likely the support for one hypothesis over another is. For instance, if $BF_{01} > 3$, the data are at least three times more likely under H_0 than under H_1. Conversely, if $BF_{01} < ⅓$, the data are at least three times more likely under H_1 than under H_0. If neither condition is met (i.e., BF's of 1/3 to 3), then the data are insufficient for distinguishing between these two models (cf. Wagenmakers et al., 2010).

In the context of assessing invariance, H_0 can be defined as the invariant model and H_1 as the non-invariant model. Similarly, this approach can be used to infer differences among single model parameters, such as, for example, differences in factor loadings or intercepts. To assess the evidence in favor of invariance, one can compute the BF_{01} for each β parameter and evaluate whether invariance is tenable ($BF_{01} > 3$), non-invariance is tenable ($BF_{01} < ⅓$), or the data are insufficient to make a decision ($⅓ < BF_{01} < 3$) (Jeffreys, 1961).

Several methods are available for approximating BFs (Kass & Raftery, 1995), and one of such is called Savage-Dickey density ratio test, which is commonly used on statistical models with equality constraints on one or more parameters. Because invariant models are nested within non-invariant models (where some difference in parameters, Δ_p, is zero), thus the BF_{01} supporting H_0 can be reduced to the ratio between the posterior probability density of H_0 and the prior belief about the state of invariance. Hence, the BF_{01} can be obtained using the Savage-Dickey density ratio: $\frac{P(\Delta_P=0|H_1,D)}{P(\Delta_P=0|H_1)}$ (Dickey, 1971; Kass & Raftery, 1995). Figure 2. 2 represents three versions of possible BF scenarios where the solid line represents the H_1 and the dotted line represents the H_0. The filled circle represents the density at zero for both hypotheses and their respective ratio corresponds to the BF. The benefit of this approach is that evidence for *both* invariance and non-invariance can be assessed for each parameter simultaneously.

Figure 2. 2

The decision rule of invariance with BFs



2.3 Method

For the current application, we first define priors for all parameters in a one factor MGCFA model with two groups. Bayesian models result in posterior distributions on parameters as functions of the data, likelihood, and priors: $p(\Lambda, \phi, \Theta, \Psi|x) \propto p(x|\Lambda, \phi, \Theta, \Psi)$ $p(\Lambda, \phi, \Theta, \Psi)$. The prior for each item i in group k can be defined as:

$$\Theta_{k,ii}^{-1} \sim \gamma(1, .5)$$

$$\mu_{ki} \sim N(0,1)$$

$$\nu_{ki} \sim N(0,1)$$

$$\Psi_k \sim N(0,1)$$

where $\Theta_{k,ii}^{-1}$ is the residual covariance, μ_{ki} is the intercept, ν_{ki} is the factor loading and Ψ_k is the latent mean.

Next, we assign $\Delta_\nu, \Delta_\tau$to denote the cross-group differences in factor loadings and item intercepts, so that the priors of their coefficients for each item i follow:

$$\beta_{\nu i} \sim N(0, \lambda_{\nu i}^2 \tau^2)$$

$$\beta_{\tau i} \sim N\left(0, \lambda_{\mu i}^2 \tau^2\right)$$

where $\lambda_i \sim C^+(0,1)$ that is commonly used in the HS prior. For the global shrinkage parameter τ, we chose a $C^+(0,1)$ that was previously proved to yield a good performance in Bayesian variable selections (Carvalho et al., 2009; Gelman, 2006).

The prior corresponding to the non-invariance hypothesis (H_1) was set to standard normal, and the posterior density at 0 was estimated using the logspline estimator (Stone, Hansen, Kooperberg, & Truong,1997). The BF_{01} was computed as the ratio of posterior and prior density at zero, for each parameter. The BF was then used to support invariance (BF_{01} > cutoffs), non-invariance (BF_{01} < 1/cutoffs), or neither due to the uncertainty in the data (Jeffreys, 1961).

2.3.1 Overview of Study 1

In order to evaluate the method of detecting non-invariant items posed above, we conducted a simulation study where a variety of conditions and datasets were generated to mimic real situations. The data generation process, the simulation design, model fitting strategy and the performance evaluation metrics are described below.

2.3.2 Data generation process

The data were simulated based on a two-group (J=2) MGCFA population model with a single latent factor, continuous items and satisfied for configural invariance. For parameters values, we followed the previous study (Shi et al., 2017). One group served as the reference group where the factor mean and factor variance are fixed to be zero and unity, respectively. While the other group served as the focal group where the factor mean and factor variance are fixed to 0.5 and 1.2, respectively. For both focal and reference group, I set the population value of all item intercepts and residual variances to 0.6 and 0.3, respectively. For the number of non-invariant items, 1/3 of items are allowed to differ both on factor loadings and intercepts (i.e.,).

2.3.3 Simulation Design

The following factors were found important to MI testing in previous studies (e.g., Liu & Aitkin, 2018; Yoon & Millsap, 2007): sample size, scale length, the amounts of non-invariant items, the magnitude of non-invariance in factor loadings, the magnitude of non-invariance in intercepts, and item reliability [1], and thus were included as simulation conditions as described below. Both groups were simulated with equal sample size including 200, 400, 600 observations per group, which was considered as small, medium, and large in

[1] To vary item reliability, the communalities of .5, .6, and .7 which represented for low, medium, and high reliabilities were used and, residual variances were set to .3, for all items.

terms of sample size. The scale length was set to either 6 items or 9 items, which there are 1/3 items were set to be non-invariant on both factor loadings and intercepts. The scale lengths were considered common for psychological measurements. The magnitude of non-invariance in factor loadings[2] were set to 0.2 and 0.4 which were considered as small to large difference, respectively. The magnitude of non-invariance in item intercepts were set to 0.15 to 0.3 which were considered as small to large difference, respectively. This led to three patterns of differences as following: 1) Small difference, where loadings and intercepts differ for 0.2 and 0.15 respectively; 2) Large difference, where loadings and intercepts differ for 0.4 and 0.3 respectively; 3) Mixed pattern one, where loadings differ for 0.4 while intercepts differ for 0.15; 4) Mixed pattern two, loadings differ for 0.2 while intercepts differ for 0.3. This yields 72 unique conditions (see Table 1), each of which was replicated 100 times.

Table 1

Simulation Design

Simulations Condition				
Sample Size	200	400	600	
Item Reliability	.45	.55	.58	
Magnitude of Difference on factor loadings	.2	.4	.4	.2
Magnitude of Difference on intercepts	.15	.3	.15	.3
Number of items	6	9		

Note. Both of the item reliability and the magnitude of difference are on a standardized scale.

2.2.4 Model fitting strategy

For the MGCFA model identification, the current study followed the reference indicator (RI) approach which had been recommended for MI testing (Rensvold & Cheung,

[2] A relative difference is used given that the differences of .2, .4 are weighted differently for different item reliabilities. More details about the equating process are available in the appendix A.

2008; Yoon & Millsap, 2007). Specifically, the mean and variance of the reference group were set to zero and one, respectively. Additionally, one invariant item was selected as RI and constrained to be equal between groups. Then, the model was fitted using Hamiltonian Monte Carlo sampling via Rstan (Stan Development Team, 2020; R Core Team, 2014) with 3000 iterations and four chains to obtain standardized posterior samples for each parameter of interest (i.e., β for all Δs). Convergence of Monte Carlo chains was assessed using the potential scale reduction factor threshold of $\hat{R} < 1.1$(Gelman & Rubin, 1992). All models estimated here converged, with all $\hat{R}$ values below 1.09.

2.2.5 Metrics of evaluating the performance of HS priors in detecting non-invariant items

Because the Bayes factor can favor invariance (H_0), non-invariance (H_1), or neither, the following five metrics were used to assess performance. The first one is *certainty*, defined as the proportion of all comparisons yielding certain decisions in the sense that the resulting BF_{01} was exceeding the threshold. Next, are *Sensitivity*, defined as the proportion of all non-invariant parameters detected as non-invariant, and *Specificity*, conversely defined as the proportion of all invariant parameters detected as invariant. These two metrics represent the ability to detect non-invariant and invariant parameters, but do not reflect accuracy. For example, specificity may be high (most invariant parameters are detected), but many non-invariant parameters may be mistakenly classified as invariant. Hence, we supplemented it with two other metrics that describe decision accuracy: *Positive predictive value* (PPV), defined as the proportion of all parameters detected as non-invariant that are truly non-invariant, and *negative* predictive value (NPV), defined as the proportion of all parameters detected as invariant that are truly invariant. In the ideal scenario, all four metrics will be 1, meaning there is a high probability of classifying parameters as non-invariant or invariant,

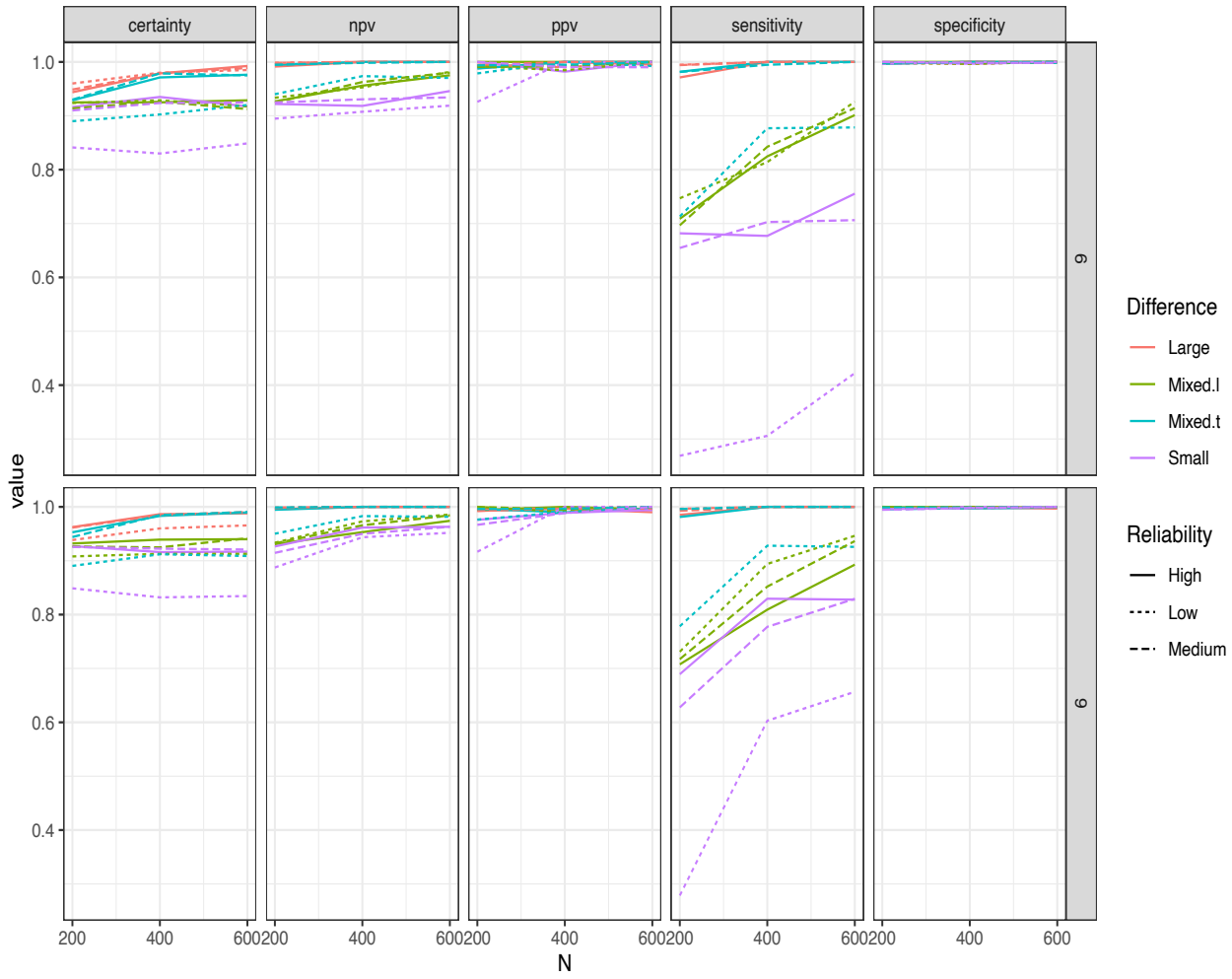
and the classification is accurate. These metrics were computed across all items on each simulation replication.

2.4 Results

The simulation results are presented in Figure 2. 3. As mentioned earlier, *certainty* describes the ability to make a decision, either for or against invariance. As expected, the magnitude of non-invariance and item reliability all showed impacts on certainty rate. In most simulation conditions, a high item reliability, and a large magnitude of non-invariance led to an increasing decision rate. Interestingly, a large sample size did not show notable impact on decision rate except when the magnitude of non-invariance was large.

Figure 2. 3

Simulation Results using BF of 3 as Decision Rule



Sensitivity describes the ability to detect non-invariance. *Sensitivity* increased with sample size across all simulation conditions. That is, as sample size increased, the probability of detecting non-invariant parameters increased as well. The magnitudes of non-invariance and item reliability altered *sensitivity*. Larger differences in parameters were more readily detectable, and *sensitivity* therefore increased, regardless of the item reliability. On the other hand, a higher item reliability increased the probability of detecting non-invariant parameters when the parameter difference was less prominent. That was to say, a smaller parameter difference was more detectable with a higher item reliability. The sensitivity rate significantly increased as a function of sample size for the 9-item scale when parameter differences were

small (see purple line in Figure 2. 3), but this was not the case for the 6-item scale. This asymmetry may be due to the reason that the number of non-invariant parameters was doubled for the 9-item scale compared to the 6-item scale, which thus increased the chance of detecting non-invariant parameters.

Specificity describes the ability to detect invariance and it was only expected to be impacted by sample size and item reliability. Notably, *specificity* was unaffected by the magnitude of non-invariance as it does not depend on parameter differences. That is, the ability to detect invariant parameters is not affected by the presence and magnitude of non-invariance in other parameters. S*pecificity* was perfect, which achieved 100% across all simulation conditions. This suggests that identifying invariant parameters was easier than non-invariant parameters using the HS prior under the current simulation conditions.

Both positive and negative predictive values were high across all conditions. This suggests that when a parameter is classified as invariant or non-invariant, the classification is generally accurate. Even in the worst case, the decision accuracy remained high on average. For example, for low reliability, small *N*, and small non-invariance conditions, predictive values were larger than .80. In the best of conditions (large sample size, reliable items, large non-invariance), the HS prior has an almost perfect probability of accurately discerning invariant from non-invariant parameters. When N was small, the result suggest that more data was needed, and the decision was neither invariance nor non-invariance.

2.5 Discussion

Measurement invariance permits meaningful comparisons in psychological constructs across different populations and measurement occasions. However, it is often found hard to achieve full MI in most empirical studies. Instead, researchers could choose alternative model

fitting strategies such as fitting a partially invariant model or deleting non-invariant items. This would require identifying all non-invariant items beforehand. Detecting non-invariant items is considered a difficult task because it can only be achieved via a series of constrained model comparisons for the most of MI assessments that focus on model-level analyses. In addition, the detection of non-invariance often relies on significance tests based on the χ^2 difference statistic or on critical differences in goodness of fit indices (e.g., Δ CFI, RMSEA; Chen, 2007; Cheung & Rensvold, 2002). These methods face two challenges: 1) Constrained model comparisons not only are cumbersome, but also inflate Type I or II rates and 2) they fail to provide direct evidence supporting non-invariance. Alternatively, Bayesian methods have shown great promise in MI assessments. For example, Verhagen et al., (2016)'s Bayes factor approach provides a method to gather supportive evidence for invariance on item-level in the IRT framework. Shi et al., (2017; 2019) utilized approximate invariance to identify all non-invariant items at once in the MGCFA framework. The current study proposed to apply a Bayesian variable selection method, where identifying non-invariant items followed a similar approach as variable selection. Specifically, we made use of the HS prior on pre-defined parameter differences for each item to compress negligible effects representing invariance while capture large effects representing non-invariance and then used Bayes factors to gather supportive evidence for making decisions.

In this dissertation, we demonstrated how the HS prior along with Bayes factors could be used to detect non-invariant items simultaneously. A Monte Carlo simulation study was conducted to examine the performance of a HS approach under several conditions varying in sample size, length of scale, item-reliability, and magnitude of parameter non-invariance. The results suggest that the HS prior variable selection approach performed well in detecting non-invariance at the item-level.

2.5.1 Considerations regarding the choice of the prior

Verhagen et al., (2016) recommended a less informative prior (i.e., Cauchy distribution) as a better option for MI testing. Shi et al., (2017) found that different prior choices did not significantly impact the MI assessments. The current study advocated to use the HS prior in detecting non-invariant items for two reasons. First, from a theoretical perspective, the prior distribution represents our pre-existing belief about invariance, which should be either yes or no before seeing the data. Indeed, the shape of HS distribution perfectly fit this situation where high density around its origin mimics the invariance and Cauchy like tail mimics the non-invariance. Second, the HS prior is known for its excellent performance in identifying effective predictors and producing BMA-like results. Yet, it should be noted that a large sample size and a large amount of difference are key to detecting non-invariance. Again, from a practical view, this situation indicates the existence of a true difference. And statistically speaking, a large sample size and magnitude of difference lead to a narrower posterior distribution which also peaks far away from zero, regardless of prior distributions.

2.5.2 Sample size considerations

Previous studies have shown that sample size is a critical factor in MI testing (Chen, 2007; Cheung & Rensvold, 2002; French & Finch, 2016). Because the χ^2 test is highly sensitive to a large sample size, a negligible model misfit may lead to the model rejection, and support for non-invariance. Unlike the NHST approach, Bayes factors evaluate the fit of *both* invariance and non-invariance hypotheses. Consequently, although negligible misfit may be present in the observed data, the invariance hypothesis may nevertheless fit better than the non-invariance hypothesis. Therefore, the effect of large sample sizes on sensitivity to negligible misfit is attenuated.

As expected, larger sample sizes produce more conclusive Bayes factors in favor of either hypothesis. When data are insufficient for deciding between non-invariance and invariance, BF's appropriately suggest that the evidence is weak. To avoid inconclusive Bayes factors, we would encourage researchers to collect sample sizes that are no less than 200 participants per group in order to accrue strong evidence either for or against measurement invariance. Moreover, the current simulation suggests that decisions based on the BF's are generally accurate, even when few decisions can be made.

2.5.3 Considerations on the threshold of BFs

The choice of BF_{01} cutoff is all arbitrary, just like the choice of an alpha level (or *p* value) in NHST (Raftery, 1995; Wagemaker, 2007). Although different in practical implication and statistical interpretation, both *p* value and BF_{01} reflect the level of researchers' confidence in making decisions. For instance, if we are confident in making a decision when one hypothesis is 3 times more likely than the competing hypothesis, choosing a relatively low Bayes factors cutoff (i.e., $BF_{01} = 3$) will be acceptable. If we think even stronger evidence is needed when making decisions (e.g., medical research, clinical trials), choosing a high Bayes factors cutoff that exceeds 20 may be more reasonable (cf. Raftery, 1995). We recommend the BF_{01} of 3 as a reasonable choice for decision making criterion in detecting non-invariant items. However, as the practical significance of statistical results may vary across research areas and from studies to studies, researchers should adjust this value accordingly.

2.5.3 Limitations and Future Directions

Despite the usefulness of the HS prior variable selection approach in detecting non-invariant items, there are several limitations to keep in mind. The current study only examined the performance of the HS prior in a simulated condition where the data were

normally distributed with no missing values, equal sample size among groups, and the model structure was simple (i.e., only contained one latent factor without item cross-loadings or correlated residual variances). For greater generalizability, future studies should investigate the performance of the HS prior in different simulation settings, such as when the data contain missing values, sample sizes are unequal across groups, or the model structure is more complex (e.g., item cross-loadings, correlated residual variance). In addition, since our study is not the only one using Bayesian method in MI testing (see Liang, & Luo, 2019; Muthén et al, 2013, 2017; Shi, Song, DiStefano, Maydeu-Olivares, McDaniel, & Jiang, 2019), future studies could compare each method and discuss some pros and cons in taking each approach.

Some studies have suggested that the traditional MI assessments allowing zero cross-group difference was too restrictive in empirical studies and thus permitted small differences in MI testing (see approximate MI, Muthen et al., 2007). In our study, the HS prior variable selection approach did not perform well either when the magnitude of non-invariance was small. Despite the sample size requirement in quantifying small differences, one should also wonder if a small difference really matters for practical purposes, such as using a composite score (assuming tau-equivalence), or factor score for the purpose of selection or diagnosis. All simulation codes are available in Appendix A (see supplemental files).

3. Chapter 3

Improving the Predictive Performance of Partially Invariant Models

3.1 The issues of failure in Measurement Invariance and potential solutions

Measurement Invariance (MI) is key to the validation of social science instruments (Lai, Richardson, & Mak, 2019; Meredith,1993). The establishment of full MI ensures the same measurement instruments can be applied with heterogenous populations (e.g., women and men), various time points, different testing formats (e.g., computerized V.S. papers), or multiple testing locations. The failure of MI may lead to a confound in the observed group differences (Lai et al., 2019; Millsap & Kwok, 2004). Nevertheless, achieving full MI is often found unrealistic in empirical studies. Indeed, often MI can only be partially fulfilled, and researchers need to search for alternative solutions. A number of strategies are available to fit the partially invariant model. One strategy is to ignore the non-invariance and still fit a fully constrained model, which has been shown to mostly produce biased results (e.g., Hisao et al., 2018; Finch & French, 2016). Another is fitting a so-called "partially constrained model", where the group constraints are relaxed on non-invariant parameters (Schmitt, Golubovich, & Leong, 2010; Shi et al., 2017). This approach usually leads to the most accurate estimates when the model is correctly specified. That is, there is no false detection of non-invariance which is not testable. A freely estimated model can also be an option that does not impose any group constraints on the model parameters (Shi et al., 2017). However, this model fitting strategy leads to an interpretational issue, because the latent construct may have different meanings between groups without equality constraints. It is also possible to delete non-invariant items when there are enough items. However, a difficult question often arises about which option should be undertaken. This is mainly because that the impact of non-invariance is understudied, such that it is unclear that whether the quantities of non-invariance (small vs.

large), and the location of non-invariance (factor loadings or intercepts) will bias the model estimates. If yes, which matters more, and what parts of model estimates these non-invariances impact the most: latent mean comparisons, regression coefficients, or correlations between factors? (e.g., Cheung & Rensvold, 2002; Lai, Richarderson & Mak, 2018; Shi et al., 2017; Vandenberg & Lance, 2000). Therefore, there is no agreement on what the best practice in terms of fitting partially invariant models should be. In other words, researchers must explore potential model solutions each time when meeting with new data. This is common practice with SEM in the frequentist domain, where the primary goal of model selection is finding a model that describes the current data best (e.g., Kaplan & Lee, 2016). However, when using measurement models for diagnostic purposes (Lai, Richarderson & Mak, 2018; Millsap & Kwok, 2004), candidate selections (Lai, Yoon & Hsiao, 2017), and assessments/evaluations (Kaplan & Lee, 2016), researchers may desire a model that not only provides good estimates for current data, but also exhibit optimal predictive performance with future observations (Kaplan & Lee, 2016). After all, the goal of scientific psychology is not only explaining the underlying mechanisms of psychological phenomenon, but also forecasting human behaviors (Fokkema, Iliescu, Greiff, & Ziegler, 2022; Yarkoni & Westfall, 2017). Nevertheless, the majority of statistical models used in the field only satisfy for explanations, and rarely get to the point of predictions. This is no exception for partially invariant models. Thus, we need to search out a prediction-focused model fitting approach for practical purpose. For examples, clinical psychologists may want to build up a model for a well-established depression diagnostic scale for future prevention purpose (Lai et al., 2018); policy makers may want to improve the predictive performance of large-scale assessments in evaluating students' progresses across-countries (Kaplan & Lee, 2017), or over time (Kaplan & Huang, 2021). Thus, this chapter will focus on investigating how to find a best model solution based on the predictive performance of measurement models when full MI fails. The

current chapter follows the following outline: 1) Presentation of the mathematical formulation of partial MI and the necessity of establishing full MI for using factor scores and measurement models, 2) discussion of current approaches using alternative model fitting strategies with partially invariant models, 3) discussion of the issue with Model selections in fitting partially invariant models and the prospective of using a Bayesian Model Averaging approach, and 4) discussion of the difficulty of averaging over the entire model space and proposition of a Horseshoe prior approach as an alternative solution.

3.1.1 Partially Invariant Models

As discussed in the previous Chapter (see Chapter 2, section 2.1.1), strong invariance is a minimal requirement for achieving estimation accuracy when the factor scores are used for mean comparisons, latent traits evaluation, or predicting external variables (Millsap, 2011; Jung & Yoon, 2016). The bias caused by the violation of MI can be illustrated in a full SEM model where an external variable is regressed on a latent factor with different populations, that is, in the case of a multigroup SEM (MGSEM) model. Following the discussion of bias in measurement and predictions by Millsap (2001), we can start with a simple linear regression containing multiple populations where the effect of group membership is not of interest. To eliminate the potential effect of group membership, a group variable is usually defined and included in the model as a predictor as follow:

$$Y = \beta_0 + \beta_1 X_{group} + \beta_2 X_{interest1} + \beta_3 X_{interest2} + \epsilon \qquad (3.1)$$

Where β_0 is the intercept, β_1, β_2 and β_3 denote the regression coefficients of X_{group} indicating group membership, $X_{interest1}$ and $X_{interest2}$, the other two key predictors of interests respectively, and ϵ is the error term. In equation (3.1), the effect of group membership on outcome Y, if there is any, will manifest in β_1, so that the estimates of β_0 β_2, β_3 and ϵ will not be contaminated. What if we ignore the effect of group membership and omit it from equation (3.1)? Specifically, let's fit a model without the grouping variable:

$$Y = \beta_0 + \beta_2 X_{interest1} + \beta_2 X_{interest2} + \epsilon \qquad (3.2)$$

when equation (3.1) reflects the actuality, so that the invisible impact of group membership will be absorbed in any of β_0, β_2, β_3 or all of them. Consequently, the estimates of β_0, β_2, β_3 will be surely biased, so is the error term. Although a simple model as equation (3.1) or (3.2) can never reflect the actuality, the difference between inclusion or exclusion of X_{group} between two models suggests that ignoring the group membership when it effectively impacts the outcome variable in regression models may lead to systematic errors on future predictions. Similarly, in MGSEM models, any violations of MI will act as significant group effects in linear regression models, so that ignoring one or more non-invariances may lead to notable estimation bias in relevant parameters and inaccurate future predictions. To be clear, let's write out the structural model concerning the regression path on an external variable Y as follows:

$$Y_k = \alpha_{0k} + \alpha_{1k}\,\Psi_{\mathrm{k}} + \zeta_k \tag{3.3}$$

where α_{0k} is an $n \times 1$ vector of intercept, α_{1k} is an $n \times 1$ vector of regression coefficient and ζ_k is an $n \times 1$ vector of residual variance with $E(\zeta_k = 0)$ and $cov(\epsilon_k, \zeta_k)$. Thus, the predicted accuracy of Y is depending on the estimate of latent factor w_k which is allowed to vary across groups for controlling the group effect. Yet, the true group effect of Ψ_{k} is related to the estimates of ν_k and β_k, and hence indirectly affected by the MI of the measurement model. As such, the attainment of strong MI is necessary in MGSEM models for achieving predictive accuracy (Millsap, 2011). For instance, one may be interested in assessing how well self-esteem predicts the quality of life while accounting for the gender difference within the given sample as well as for the future observations. Without knowing the invariance status of the measurement structure of self-esteem between men and women, one cannot say for sure that the gender effect is truly being accounted. The gender effect on predictors could be contaminated by the gender difference in the measurement model. Nevertheless, fulfilling all levels of invariance is known as being empirically difficult and some of the non-invariances are often observed in ν_k or β_k, or both. Thus, partial MI is considered as being more realistic for empirical studies (Lai et al., 2017; Millsap et al., 2004). With partially invariant models, controlling for group effects means allowing some levels of measurement non-invariance. If non-invariances are ignored and equal group constraints are enforced (e.g., fitting a strong invariant model) to maintain MI, the prediction of Y should be expected low in accuracy. Although there is no cure for the failure of MI, some works exploring different model fitting strategies still have been done in an effort to utilize partially invariant models (e.g., Shi et al., 2017; Hsiao & Lai, 2018; Lai, Richardson & Mak, 2018; Lai et al., 2017). In

the next section, I will discuss current approaches in using different model fitting strategies with partially invariant models in a great detail.

3.1.2 Current approaches in fitting models with partial MI

MI testing aims for two goals: 1) validating the measurement structure of scales/tests, 2) making sure the measurement model can be used as a part of full SEM or comparable across groups. In the former case, the task is completed either when full MI is established, or any of the non-invariant items are identified. Things start to become complicated in the latter case. When full MI is established, the task is completed given that either mean comparisons or full SEM can be proceeded with the measurement model. When full MI fails, researchers now have two options: 1) End the study and claim that MI fails so that it is invalid to use the measurement with heterogenous populations, or 2) identify the causes of the measurement non-invariance and try to fit a partially invariant model. The second option is often preferred in practice. Therefore, researchers often turn to search for alternative strategies to fit partially invariant models.

However, as discussed previously, finding a good solution for partially invariant model is never an easy task. Challenges arise to fit a partially invariant model mainly based on two considerations. On one hand, we want the latent construct between groups to remain theoretically unchanged. Yet, without imposing full equality across groups, group means are difficult to interpret. On the other hand, the potential estimation bias on factor mean comparisons, path coefficients, or cross-factor correlations of different model fitting strategies may vary case by case, since it could be influenced by the number and quantity of non-invariance which are usually "unknown". The amount of bias in parameter estimates caused by different model fitting strategies is subject to a set of factors such as sample size, the pattern and magnitude of non-invariance, and scale length (e.g., Byrne et al., 1989;

Schmitt & Kuljanin, 2008; Shi et al., 2017). For instance, when the magnitude of non-invariance is small, the difference in estimation bias among different model fitting strategies (e.g., fully constrained model versus partially constrained model) may be quite trivial (Shi et al., 2019); when the magnitude of non-invariance is small, deleting the non-invariant item or releasing the group constraint on non-invariant parameter may not lead to a much better result compared to fitting a fully constrained model. As such, some of these model fitting strategies may be more preferable than the others under certain conditions (Shi et al., 2019) (e.g., better model fit, higher more predictive accuracy, more theoretically sound).

Several studies have been done that explore the impact of implementing different model fitting strategies with partially invariant models, such as examining the selection accuracy (Lai, Richardson & Mak, 2018; Lai, Kwok, Yoon & Hsiao, 2017; Millsap & Kwok, 2004; Hisao & Lai, 2018), comparing the estimation bias in factor covariances, mean structures, regression coefficients and moderation effects (Byrne, Shavelson & Muthen, 1989; Shi et al., 2019). This body of research underscores the practical significance of non-invariance. They suggested that researchers should not stop at where full MI fails. Instead, other options should be explored for empirical uses, such as fitting partially invariant models, or using composite scores if the amount and of bias are acceptable. For instance, given a partially invariant model, how wrong could a decision on evaluation or selection be if a composite/sum score (Lai et al., 2017, 2018; Millsap & Kwok, 2004); how different partially invariant model fitting strategies will affect the estimates of other model parameters (Byrne, Shavelson & Muthen, 1989; Shi et al., 2019; Millsap, 2011; Hisao & Lai, 2018). Another line of research focuses on the practical inferences of using partially invariant models for predictions. Millsap (2011) has shown that the failure of attaining MI in exogenous variables may lead to notable estimation bias on the intercept of endogenous variables, but not so much

on the regression coefficient. Hisao & Lai (2018) investigated the impact of partial invariance on the moderation effects with multilevel data. This line of studies is relatively sparse and has not yet provided any conclusive answers regarding which model fitting strategies should be taken over others.

3.1.3 Model Averaging Versus Model Selection

Although identifying the best model is always desirable, there are some situations where different model fitting strategies are indistinguishable in terms of model fit or producing bias (e.g., produce similar estimates, fit data equally well). More importantly, no matter which model fitting strategy is selected, we are making inferences about population conditional on one single model. The fundamental issue of making decisions from a single model is that it ignores the uncertainty in model selection (e.g., Kaplan & Lee, 2018; Madigan & Raftery, 2012). The risk of being overly confident in the inference and decisions made from one single model is often underestimated (Hoeting, Madigan, Raftery & Volinsky, 1999). This is especially the case for partially invariant models, where the pattern of non-invariance can be complex and often varies case by case (Lai et al., 2017). Thus, taking a single model fitting strategy (e.g., fitting a partially constrained model) may not be optimal for estimating other parameters of interest, more importantly, for making scientific inference about population in general. Alternatively, there is a growing interest in employing Bayesian Model Averaging (BMA) techniques in SEM to improve models' predictive performance and solving the uncertainty arising from model selections (Kaplan & Lee, 2016; Raftery, 1992).

As previously discussed, many efforts have been made to search for the best model fitting strategy when MGCFA only holds for partial invariance. There are several limitations with existing methods. First, there are several factors affecting the final decision in selecting the best model fitting strategy (e.g., a fully constrained model, deleting non-invariant items, a

partially invariant model, a freely estimated models) such as the model selection criteria (e.g., model fit, predictive accuracy, estimation bias), the length of scale, sample size, or the amount of non-invariance (i.e., numbers, magnitudes). Second, the candidate models set increases as the non-invariant situations become more complex. For instance, the magnitude of non-invariance may be negligible on some parameters, but notable on others; some items may only establish for loading invariance, but not intercepts invariance. This leads to a large variation among different model fitting strategies within the candidate models set. Therefore, one problem arises during this searching procedure, that is, no matter which model is being selected, it is still questionable whether we can generalize the final solution to the general population. The popularity in model selection suggests that people often overlook the model uncertainty resulting from the large variation between potential model fitting strategies and be overly confident about the single model inference (Hoeting et al, 1999). To address this issue, the current study proposes a Bayesian model averaging (BMA) approach to accounting for the uncertainties in model selection procedure with partially invariant models.

The BMA approach incorporates the model uncertainties by utilizing the information from all potential models (Clyde & Iversen, 2013; Hoeting et al, 1999; Madigan & Raftery, 1994). Specifically, BMA estimates all candidate models to obtain averaged parameter estimates which are weighted by each model's corresponding posterior probability. Several studies suggest that BMA provides a better out of sample prediction compared to any single model solution (Hoeting et al., 1999; Madigan et al., 1994). Kaplan et al., (2016; 2018) have discussed the application of BMA in SEM and showed that BMA exhibited a good predictive performance. According to Kaplan et al., (2016; 2018), BMA is particularly suitable for SEM when there are some competing theories about whether some paths should be included/excluded, and when researchers aim for improving model predictions. The same

logic can be applied for improving the prediction of partially invariant models, where the numbers and locations of imposing/relaxing equal group constraints are often uncertain. Therefore, we argue that using BMA technique could be a better solution rather than selecting a single best model.

3.1.4 Bayesian Model Averaging

A combined model approach using the Bayesian method has received extensive attention in statistical literature over the last decade (Clyde & Iversen, 2013). Two different frameworks of BMA have been discussed in the statistical literature (Bernado & Smith, 1994; Clyde & Iversen, 2013; Hoeting et al., 1999; Navarro, 2018; Vehtari, Simpson, Yao & Gelman, 2019). One view is referred as *M-closed,* where one holds belief that the true data generating model M_t is unknown but is included in a set of candidate models $M = \{M_j, j = 1 \ldots J\}$(e.g., Hoeting et al., 1999). The other perspective is referred as *M-open*, where M_t is no longer a part of $\{M_J\}$, but can be approximated by using information provided by $\{M_J\}$ (Bernado & Smith, 1994; Clyde & Iversen, 2013). For current application, we will focus on M-closed framework and assume that the true data generating process is a part of $\{M_J\}$.

In M-closed framework, the true data generating model M_t is included in the set of candidate models where $M = \{M_j, j = 1, 2, \ldots J\}$. Recall that the quantity of interest in current study is cross-group difference for each parameter of interest (e.g., factor loadings, intercepts), which is defined as Δ. Hence its posterior distribution conditional on a dataset D can then be expressed as:

$$pr(D) = \Sigma_{j=1}^{J} pr\big(M_j, D\big) pr(D) \tag{3.4}$$

which is the averaged posterior distribution of Δ under each potential model in $\{M_J\}$ that is weighted by their posterior model probability, which can be written as:

$$pr(M_j \mid D) = \frac{pr(D \mid M_j)pr(M_j)}{\Sigma_{j=1}^{J} pr(D \mid M_j)pr(M_j)} \tag{3.5}$$

where

$$pr(D \mid M_j) = \int pr(D \mid \theta_j, M_j)pr(\theta_j \mid M_j)d\theta_j \tag{3.6}$$

is the marginal likelihood of M_j. θ_j refers to the vectors of parameters in M_j (e.g., for CFA model, $\theta_j = (\Lambda, \tau, \Phi, \Theta, \alpha)$), $pr(\theta_j \mid M_j)$ denotes the prior density of θ_j under M_j, $pr(D \mid \theta_j, M_j)$ is the likelihood and $pr(M_j)$ stands for the prior probability that M_j is M_t. The posterior means and variance for Δ can be expressed as:

$$E[\Delta \mid D] = \sum_{j=1}^{J} E(\Delta \mid M_j, D)pr(M_j \mid D) = \sum_{j=1}^{J} pr(M_j \mid D)\widehat{\Delta_j} \tag{3.7}$$

$$Var[D] = \sum_{j=1}^{J} (Var[\Delta \mid D, M_j] + \widehat{\Delta_j}^2)pr(M_j \mid D) - E[\Delta \mid D]^2 \tag{3.8}$$

This approach gets to the idea that the true model is within a known model searching space. Thus, M_t will likely get the largest weight, while the model that differs from M_t (contains less useful information) will get relatively smaller weight. Compared to any single model from $\{M_J\}$, the averaged model solution has the advantage of covering both major information from the true model given the current data, and the peripheral information from

other potential models which may be ture in the general population. Hence, BMA is expected to provide a better predictive accuracy for future observations which has been shown via a logarithmic scoring rule:

$$-E\ [log\ \{\sum_{j=1}^{J} pr(\Delta \mid M_j, D)pr(M_j|D)\ \}]\ \leq\ -E[log\ \{pr(\Delta|\ M_j, D)\}]$$

(Madigan et al., 1994, 1997). According to this logarithmic scoring rule, the smaller the value of the score is, the better the predictive task has performed (Good, 1995).

3.1.5 The difficulty in BMA and using a Horseshoe Prior as an equivalent solution

The essential issue of applying BMA techniques is that we have to build each possible model for constructing a candidate models set, and in a large SEM model this can result in an enormous model space (Hoeting, Madigan, Raftery, & Volinsky, 1999; Kaplan & Lee, 2016). In our special case -- partially invariant models, we can start with three models: M_0: a fully constrained model, M_n: a freely estimated model, M_i:a partially constrained model as "anchoring" models. We view the model space $\{M_J\}$ as a continuum with limits, and then set M_0 as its origin, M_n as its endpoint, and M_i as a mid-point that locates on this continuum somewhere between M_0 and M_n. The size of $\{M_J\}$ can be small if one limits the uncertainty only among anchoring models. Then we can obtain a size of $\{M_J\}$ which renders the summation of equation (3.7, 3.8) manageable. Yet, if one acknowledges that there are still some uncertainties among these anchoring models along the continuum, the size of $\{M_J\}$ can grow much bigger. For instance, some non-invariant parameters may contain less group information (i.e., the magnitude of non-invariance is smaller), while the others contain more (i.e., the magnitude of non-invariance is more). Constraining those parameters with smaller amounts of non-invariance may not lead to a very different model. Similarly, deleting these

non-invariant items with low reliability (i.e., smaller factor loading) may not be the same compared to deleting highly reliable items. If we take all these potential models into considerations, the size of $\{M_J\}$ can grow to infinity, which leads to the complete summation of equation (3.7, 3.8) impractical. Two methods have been discussed in the literature to address this computational issue. The first one is called "Occam window" which attempts to reduce the size of $\{M_J\}$ by excluding some trivial models (e.g., Madigan & Raftery, 1994; Kaplan & Lee, 2018), and only use the models that are retained by "Occam razor" to get the weighted average posterior distribution of the parameters of interest. The second approach lies in placing the horseshoe (HS) prior on the parameter of interests/uncertainties which has been shown producing "BMA-like" result and does not require collecting all potential models as the "Occam window", hence is much less computationally demanding (Carvalho, Polson, & Scott, 2010; Piironen & Vehtari, 2016).

Of course, the HS prior cannot fully substitute BMA under all conditions, but it can be used on partially invariant models: The main reason is that on the continuum of $\{M_J\}$, all possible models are nested within each other such that one model can be turned into another by simply omitting/adding a path. This "turning on/off" action can be easily handled by placing the HS prior on the path of particular interest (e.g., cross group difference on loadings). Therefore, for improving the prediction of partially invariant models, we propose to use the HS prior as an equivalent solution to BMA. In the section 3.2, we will provide technical details about how to use the HS priors to with partially invariant models.

3.2 Using the Horseshoe Prior to Improve the Prediction of Partially Invariant Models

3.2.1 The Horseshoe Prior

The HS prior is frequently used to solve variable selection problem in Bayesian sparse learning literature and has been proven to produce BMA like estimates (e.g., Carvalho, Polson, & Scott, 2010; Piironen & Vehtari, 2016). Let's start by introducing the HS prior in a simple example with linear regression where $(\beta) \sim N(\beta, \sigma^2 I)$ and effects β of predictors are assumed to be sparse. The HS prior states that each β_i holds for conditional independence with a density of $\pi(\tau)$ which can be written as a scale mixture distribution (see Chapter 2, section 2.2.4, equation 2.1). When handling sparsity issues, the global shrinkage parameter τ regulates all β towards zero, while the local shrinkage parameter λ_i permits some effective β_i to escape from the shrinkage. (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2016, 2017).

The statistical property of the HS prior meets the model assumption of MI where there is a small number of items that are expected to be non-invariant. In other words, some β values should deviate significantly from zero and take on meaningful quantities, while other β values, which are around zero, should be compressed towards zero. To illustrate, we borrow the idea from Carvalho et al., (2010), firstly assume that $\tau^2 = 1$ and define a random shrinkage weight $k_i = 1/(1 + \lambda_i^2)$, which stands for the amount of weights that the posterior means of β_i takes on zero given the data y (see Chapter 2, section 2.2.4, equation 2.2, 2.3). Given the prior of k_i depends on the different $\pi(\lambda_i)$, along with this transformation, we can then get a clear idea about the advantage of applying the HS prior to distinguish between non-invariance (signals) and invariance (noises). As in equation (2.1), $\lambda_i \sim C^+(0,1)$ suggests that $k_i \sim Be(1/2,1/2)$, a symmetric density that is bounded between 0 and 1. This horseshoe shaped distribution indicates that two things are expected in the data: Strong signals indicate

non-invariances ($k \approx 0$, no shrikage), and noises indicate invariances ($k \approx 1$, complete shrinkage).

3.2.2 The Role of the Horseshoe Prior in Partially Invariant Models

To mimic the performance of BMA, the current approach uses the HS prior to incorporate unknown cross-group non-invariance (e.g., numbers, magnitude) into measurement models. Specifically, the HS prior works as a parameter "switch" in MGCFA models given its statistical property (see *Figure 2. 1*) (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2016, 2017).

Figure 2. 1, so that the switch turns off when the magnitude of non-invariances is large enough, while turns on when the magnitude of non-invariance is ignorable. Therefore, the notable non-invariance will be automatically taken as signal/effect and estimated. While the small non-invariance will be recognized as noise and regularized to zero. This approach not only improves the model's prediction, but also simplifies the model fitting procedure. The traditional method usually accomplishes the model fitting procedure in three steps, 1) identifying non-invariant items/parameters, 2) picking the most appropriate model fitting strategies accordingly, 3) fitting the "selected" model to the data. By contrast, the HS prior approach combines the detecting the non-invariance (see Chapter two for details) and selecting & fitting the model into one step.

To implement the HS prior as a parameter "switch", a set of hyperparameters Δs standing for the cross-group difference of each item in factor loading (Δ_λ), item intercepts (Δ_τ), and as well as residual variances (Δ_ϵ) are defined. The HS prior is then placed on these Δs to handle these undetermined noninvariances. As discussed above, the excellent performance of the HS prior in solving sparse problems makes it suitable for any given

situations where signals/effective predictors are believed to be scarce. For well-established measurements, the majority of scale items should be expected to hold for invariance. That is, only few of Δs should be expected to be relevant/significant. The HS prior's tall spike around the origin shrinks insignificant Δs (invariant parameters) towards an infinitely small value, while its flat Cauchy-like tail allows relevant Δs (non-invariant parameters) remain large.

In this section, it has been demonstrated that the HS prior would be a good substitution for BMA for improving the prediction of partially invariant model. The section follows moves on to consider the choice of model selection/comparison tool in assessing the predictive ability of the HS prior model.

3.2.3 Measures of Predictive Ability of Partially Invariant Models with the Horseshoe Prior

There are two classes of techniques in measuring models' out-of-sample predictive accuracy: cross-validation and information criteria (see a review, Kelter, 2021; Vehtari, Gelman, & Gabry, 2017). For the current approach, we consider a log-score based measure – expected logarithm pointwise predictive density (elpd) via leave-one-out cross validation as the best method, since it tends to select the model that generates highest probability for new data (Vehtari, Gelman, & Gabry, 2017).

Let's consider n observations $y_1, y_2 \dots y_n$ given parameters θ, and hence $p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$. This formulation can also include latent variables f_i so that $p(y|f,\theta) = \prod_{i=1}^{n} p(y_i|f_i\theta)$. With a prior distribution $p(\theta)$, we could obtain a posterior distribution $p(\theta|y)$ and a posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y_i}|\theta)p(\theta|y)\, d\theta$. Hence, the elpd can be written as:

$$\sum_{i=1}^{n} \int p_t(\tilde{y}_i) log p\,(\tilde{y}_i|y)\, d\tilde{y}_i \tag{3.9}$$

Where $p_t(\tilde{y}_i)$ refers to the distribution of true data-generating process for $\tilde{y}_i$. The elpd is usually approximated via the Bayesian leave-one-out cross-validation (LOO-CV) given the true data-generating process is unknown, which follows:

$$elpd_{loo} = \sum_{i=1}^{n} log\, p(y_i|y_{-i}) \tag{3.10}$$

Where $p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})\, d\theta$ is the LOO predictive density given the data without the *i*th data point (Vehtari, Gelman, & Gabry, 2017). To reduce the computational difficulty and maintain the estimation quality, we used Pareto smoothed importance sampling (PSIS, Vehtari, Simpson, Gelman, Yao & Gabry, 2015) for estimating elpd-loo as shown below:

$$\widehat{elpd_{psis-loo}} = \sum_{i=1}^{n} log\left(\frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s}\right) \tag{3.11}$$

Where *s* denotes the number of posterior simulations, *w* denotes the truncated importance sampling weights.

When assessing the expected predictive accuracy of a single model, the higher values of $\widehat{elpd_{loo}}$ suggest a better out-of-sample predictive accuracy (Bürkner & Vuorre, 2018; Vehtari et al., 2017). When comparing the expected predictive accuracy between two fitted models, we can estimate the difference in $\widehat{elpd_{loo}}$ so as its corresponding standard error. Specifically, assuming model A is compared with model B, with corresponding $\widehat{elpd_{loo-A}} =$

$\sum_{i=1}^{n} \widehat{elpd_{loo-A,i}}$ and $\widehat{elpd_{loo-B}} = \sum_{i=1}^{n} \widehat{elpd_{loo-B,i}}$, where n is the number of independent sample draw. Thus, the difference in their expected predictive accuracy is $\Delta\widehat{elpd_{loo-AB}} = \widehat{elpd_{loo-A}} - \widehat{elpd_{loo-B}}$, and the corresponding se of the difference can be computed as: $se(\Delta\widehat{elpd_{loo-AB}}) = sd_{i=1}^{n}(\widehat{elpd_{loo-A,i}} - \widehat{elpd_{loo-B,i}}) * \sqrt{n}$. A ratio between the difference in $\widehat{elpd_{loo}}$ and the se of the difference: $\frac{\Delta\widehat{elpd_{loo-AB}}}{se(\Delta\widehat{elpd_{loo-AB}})}$ is used to select a model with better performance. A rule of thumb $\frac{\Delta\widehat{elpd_{loo-AB}}}{se(\Delta\widehat{elpd_{loo-AB}})} > 2$ indicates that a model performed significantly better than the other in terms of out-of-sample prediction (Bürkner, & Vuorre, 2018; Vehtari, Gelman, & Gabry, 2017).

3.3 Method

For the current simulation study, a population MGSEM model is defined where a manifest variable is regressed on a single latent factor with the varying number of items between two groups. Four model fitting strategies are adopted to accommodate the MGSEM model that only holds for partial invariance as following: 1) a freely estimated MGSEM where no equal group constraint is imposed; 2) a constrained MGSEM model where the equal group constraint is imposed on each parameter that is associated with the measurement model ; 3) a partially constrained MGSEM model where the equal group constraint is only imposed on the invariant parameter; 4) a HS constrained MGSEM where a set of predefined Δ parameters representing parameter difference were placed with HS priors. All these models are fitted using Bayesian estimations where the priors and posterior are described in the following section.

We first define priors for all parameters as following:

$$\Theta_{k,ii}^{-1} \sim \gamma(1, .5)$$

$$\mu_{ki} \sim N(0,1)$$

$$\nu_{ki} \sim N(0,1)$$

$$\Psi_k \sim N(0,1)$$

$$\alpha_{0k} \sim N(0,1)$$

$$\alpha_{1k} \sim N(0,1)$$

$$\zeta_k \sim \gamma(1,.5)$$

where i refers to the number of items in the measurement model, k refers to the group membership. Hence, $\Theta_{k,ii}^{-1}$ is the residual covariance among items, μ_{ki} is the item intercept, ν_{ki} is the item loading and Ψ_k is the latent mean. For the s α_{0k} is the constant, α_{1k} is the regression coefficient and ζ_k is the residual variance.

Next, we assign Δ_ν, Δ_τ to denote the cross-group differences in factor loadings and item intercepts, so that the priors of their coefficients for each item i follow:

$$\beta_{\nu i} \sim N(0, \lambda_{\nu i}^2 \tau^2)$$

$$\beta_{\tau i} \sim N\left(0, \lambda_{\mu i}^2 \tau^2\right)$$

where $\lambda_i \sim C^+(0,1)$ that is commonly used in the HS prior. For the global shrinkage parameter τ, we chose a $C^+(0,1)$ that was previously proved to yield a good performance in Bayesian variable selections (Carvalho et al., 2009; Gelman, 2006). Therefore, the resulted

posterior distributions on parameters as functions of the data, likelihood, and priors: $p(\Lambda, \phi, \Theta, \Psi, \alpha_0, \alpha_1, \zeta | x) \propto p(x|\Lambda, \phi, \Theta, \Psi, \alpha_0, \alpha_1, \zeta)\, p(\Lambda, \phi, \Theta, \Psi, \alpha_0, \alpha_1, \zeta)$.

3.3.1 Overview of Study 2

For evaluating the predictive abilities of MGSEM models using the HS prior, we conducted a simulation study where a variety of conditions and datasets were generated to be representative of a real situation. Four model fitting strategies mentioned above were compared in terms of out-of-sample prediction via the efficient approximate leave-one-out cross-validation (Vehtari, Gabry, Magnusson, Yao, Bürkner, Paananen, Gelman, 2020). The data generation process, simulation design, model fitting strategies and the performance evaluation metrics are described as below.

3.3.2 Data generation process

The data were simulated based on a two-group (J=2) MGSEM population model with a single latent factor which contains continuous items and holds for configural invariance. One group is served as the reference group where the factor mean and factor variance are fixed to be zero and unity, respectively. The other group serves as the focal group where the factor mean and factor variance are fixed to 0.5 and 1.2, respectively. For both focal and reference group, we set the population value of all item intercepts and residual variances to 0.6 and 0.3, respectively. For the number of non-invariant items, 1/3 of items are allowed to differ both on factor loadings and intercepts. For the structural part, the focal group was set with a standardized regression coefficient of .45 and a constant of .5, while the reference group was set with a regression coefficient of .4 and a constant of .2. And the residual variance was set to .3 for both groups.

3.3.3 Simulation Design

The simulation design is the same with Study 1 (see Chapter 2, Table 1).

3.3.4 Model fitting strategy

Recall that four models fitting strategies are compared : 1) a freely estimated MGSEM where no equal group constrain is imposed; 2) a constrained MGSEM model where the equal group constrain is imposed on each parameter that is associated with the measurement model ;3) a partially constrained MGSEM model where the equal group constrain is only imposed on the invariant parameter; 4) a HS constrained MGSEM where a set of predefined Δ parameters representing parameter difference were placed with the HS prior. For the measurement model identification, the current study followed the reference indicator (RI) approach which had been recommended for fitting MGSEM models (Rensvold & Cheung, 2008; Yoon & Millsap, 2007). Specifically, the mean and variance of the reference group were set to zero and one, respectively. Additionally, one invariant item was selected as RI and constrained to be equal between groups. Then, each model was fitted using the Stan sampler from Rstan with 3000 iterations and four chains and its corresponding $\widehat{elpd_{loo}}$ is estimated with 2000 sample draws using "loo" package (Stan Development Team, 2020; R Core Team, 2014; Vehtari, Gabry, Magnusson, Yao, Bürkner, Paananen, Gelman, 2020). Convergence of Monte Carlo chains was assessed using the potential scale reduction factor threshold of $\hat{R} < 1.1$ (Gelman & Rubin, 1992). All models estimated here converged, with all $\hat{R}$ values below 1.09.

3.3.5 Metrics of evaluating the predictive performance of the HS prior in Multi-group SEM

One metric is used to evaluate the predictive performance of different model fitting strategies. We first focus on the difference in $\widehat{elpd_{psis-loo}}$ between two models, where a higher value of $\widehat{elpd_{psis-loo}}$ suggests a better model performance. Then, we consider the ratio between the difference in $\widehat{elpd_{psis-loo}}$ and the standard error of the difference in $\widehat{elpd_{psis-loo}}$, where a ratio above 2 suggests that the performance of one model is significantly better than the other. As such, we created one metric: *absolute best*. For each iteration, the model will be scored 1 on *absolute best* if it is significantly better than other models according to the rule of thumb: $\frac{\Delta\widehat{elpd_{psis-loo-AB}}}{se(\Delta\widehat{elpd_{psis-loo-AB}})} > 2$ (Bürkner, & Vuorre, 2018; Vehtari, Gelman, & Gabry, 2017). It should be noted there may be no model scored 1 on *absolute best* under some conditions where the difference between model with the highest $\widehat{elpd_{psis-loo}}$ and the model with the second highest $\widehat{elpd_{psis-loo}}$ is trivial.

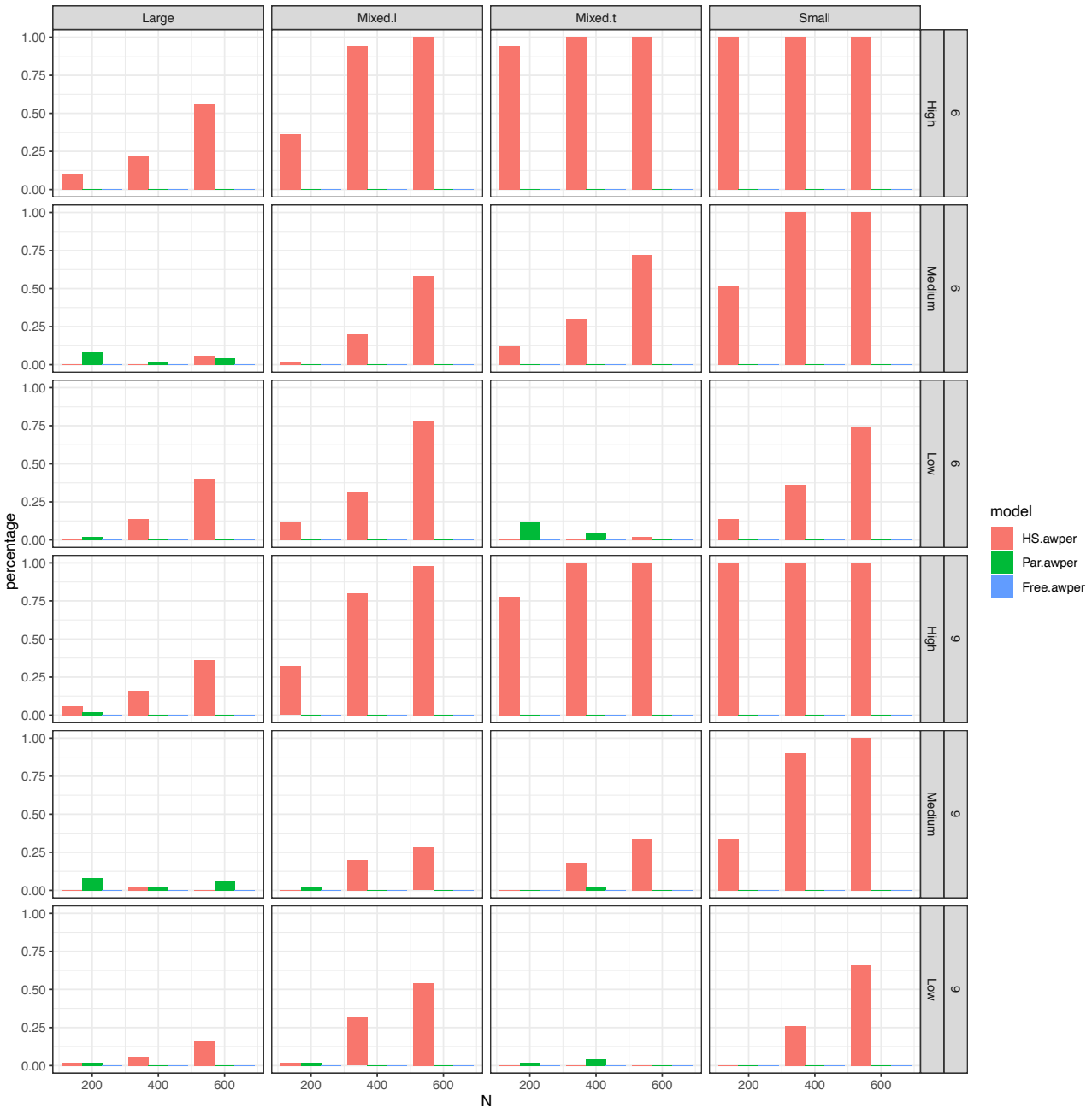
3.4 Results

The simulation results are presented in

Figure 3. 1. As mentioned earlier *absolute best* rate describes the condition where the *absolute best* model is significantly better than all other models according to the rule of thumb $\frac{\Delta elpd_{\widehat{psis-loo-AB}}}{se(\Delta elpd_{\widehat{psis-loo-AB}})} > 2$. Therefore, the rate of *absolute best* is the percentage of one model scored 1 on *absolute best* over all iterations for each condition. In accordance with *absolute best* rate, the constrained model has the lowest $\widehat{elpd_{psis-loo}}$ throughout all simulation conditions and the HS prior model has the highest $\widehat{elpd_{psis-loo}}$ for most of the simulation conditions. The magnitude of non-invariance and item reliability all showed impacts on the *absolute best* rate of the HS prior model. In most simulation conditions, a large sample size, a high item reliability, and a small magnitude of non-invariance led to a higher *absolute best* rate of the HS prior model. In contrast, when the sample size was small, the item reliability was medium or low, and the magnitude of non-invariance was large, the performances of the HS prior model and the partially constrained model were indistinguishable. Also, there was no notable difference in the *absolute best* rate of the HS prior model between the 6-item scale and the 9-item scale.

Figure 3. 1

The percentage of absolute best rate over 100 simulations


Large
Mixed.l
Mixed.t
Small
High
Medium
Low
6
9
percentage
1.00
0.75
0.50
0.25
0.00
200
400
600
N
model
HS.awper
Par.awper
Free.awper

3.5 Discussion

When full MI fails, researchers often turn to searching for alternative model fitting strategies to fit partially invariant models. The previous studies mostly focused on finding a model solution that yields the best model fit or produces minimal estimation bias (e.g., Hisao et al., 2018; Shi et al., 2019). While there was a lack of research concerning the practical inference of fitting partially invariant models (Lai et al., 2017), that is, finding a model that not only fits current data well, but also provides a good estimate for future observations or other populations (Kelter, 2021). In addition, some potential model fitting strategies such as a partially constrained model, or a fully constrained with non-invariant item deleted, would require identifying all non-invariant items beforehand. However, the detection of non-invariant items is considered a difficult task (Chen, 2007; Cheung & Rensvold, 2002) and its accuracy cannot be examined. Also, it is hard to decide on a single model because different models could perform equally well under certain conditions. More importantly, the inference generated from a single model solution which does not account for the model space uncertainty may be lack of generalizability. Instead, the BMA technique could be used to accommodate partially invariant models since it accounts for uncertainties in the model space and provides a good out-of-sample prediction. Given the known computational difficulty in BMA, we used the HS prior as a substitution. Specifically, we applied the HS prior as a "switch" on the pre-defined parameter differences of each item to mimic BMA-like performance. The current study showed that the predictive ability of partially invariant models could be improved by the HS prior.

In this chapter, we demonstrated how to use the HS prior to fit partially invariant MGSEM models without detecting non-invariant items. A Monte Carlo simulation study was conducted to evaluate the performance of a HS prior model in comparison with other models

under several conditions varying in sample size, length of scale, item-reliability, and magnitude of parameter non-invariance. According to simulation results, the HS prior model showed better out-of-sample predictions over other models under different conditions.

3.5.1 Sample size consideration

Previous studies have shown that the sample size is a critical factor in assessing predictive performance of BMA-SEM (Kaplan & Lee, 2016) and the HS prior model (Li, Craig & Bhadra, 2019). The predictive performance of BMA-SEM and the HS prior model benefits from a large sample size. Specifically, BMA-SEM showed better predictive performance in comparison with other approaches (e.g., Bayesian SEM) and the HS prior model produced lower bias on non-zero elements when the sample size is sufficiently large. Similarly, our results showed that as the sample size increased, the predictive performance of the HS prior model became distinguishably better than other models. However, a small sample size should not be a concern for using the HS prior model. As shown in

Figure 3. 1, although the partially constrained model produced a higher $\widehat{elpd_{psis-loo}}$ with a small sample size than the HS prior model, the difference was negligible.

3.5.2 Equivalent Performance between the HS prior model and partially constrained model

Previous studies have shown that partially constrained models often led to the most accurate parameter estimates and better model fit (Hisao et al., 2018; Shi et al., 2019). Our results showed that under certain conditions, the partially invariant model performed as well as the HS prior model. This result is not surprising since the partially invariant model representing the true data generating process and should be performing well even when the information is sparse (e.g., small sample size, low reliability). However, this can only happen in simulations since the truly non-invariant items/parameters are known. In practice, fitting a partially constrained model would require the detection of non-invariance beforehand which cannot be guaranteed for accuracy.

3.5.3 Limitations and Future Directions

Despite the good performance of the HS prior in fitting partially invariant models, there are several limitations to keep in mind. The current study only examined the performance of the HS prior in a simulated condition where the data were normally

distributed with no missing values, equal sample size among groups, and the model structure was simple (i.e., only contained one latent factor without item cross-loadings or correlated residual variances and the dependent variable was manifest). For greater generalizability, future studies should investigate the performance of the HS prior in different simulation settings, such as when the data contain missing values, sample sizes are unequal across groups, or the model structure is more complex (e.g., item cross-loadings, correlated residual variance, a full SEM model). Also, the current study only compared the HS prior model with three commonly used models, future studies could include other models such as composite score models, fully constrained model without non-invariant items. Additionally, since our study is not the only one using Bayesian method to incorporate non-invariance into measurement model (see approximate Bayesian MI in Liang, & Luo, 2019; Muthén et al, 2013, 2017), future study could compare each method and discuss some pros & cons in taking each approach.

As discussed in the previous chapter, whether the equality-constraint between groups is too restrictive, or what amount of cross-group non-invariance that should be allowed is often under debate (see approximate MI, Muthen et al., 2007). In our study, the quantity of non-invariance is auto adjusted by the HS prior given data. Other regularization priors such as Lasso, Spike and Slab may be also useful in calibrating the quantity of non-invariance that should be allowed. Future studies should compare the impact of using different regularization priors in quantifying permittable non-invariance on the predictive performance of partially invariant models. All simulation codes are available in Appendix B (see supplemental files).

4. Chapter 4

Empirical Analysis of Measurement Models with Horseshoe Priors

4.1 Introduction

In the previous two chapters, we have demonstrated the advantages of employing the horseshoe (HS) prior in studying measurement invariance (MI) via simulation studies. Although simulation is a great way to help us identifying some crucial factors when developing statistical methods, the real situation can never be fully mimicked. Besides, the purpose of improving existing methods and developing new methods is to assist empirical studies. Thus, this chapter aims to show a general audience how to implement the HS prior in studying MI using data from two empirical studies.

In the first demonstration, we will show how to detect non-invariant items using the HS prior along with Bayes Factors (BF). In the second demonstration, we will show how to incorporate non-invariance into a partially invariant multigroup structural equation model (MGSEM) via the HS prior and illustrate the advantage of the predictive performance of the HS prior model.

4.2 Evaluating the item equality of CES-D between groups

Here we provide two empirical analyses to illustrate the application of the HS prior in assessing MI where 1) longitudinal MI was evaluated, 2) multi-group MI was evaluated. In both examples, we evaluated the MI on participants' depressive symptoms across two measurement occasions and between gender. We used the data from the Health and Retirement Study (HRS), where the latent construct of depressive symptoms among older adults was measured by the 9-item version of the Center for Epidemiological Studies Depression Scale (CES-D) (Kohout, Berkman, Evans, & Cornoni-Huntley, 1993; Radloff, 1977).

For the current application, we first defined priors for all parameters in a one factor MGCFA model with two groups. Bayesian models resulted in posterior distributions on parameters as functions of the data, likelihood, and priors: $p(\Lambda, \phi, \Theta, \Psi|x) \propto p(x|\Lambda, \phi, \Theta, \Psi)$ $p(\Lambda, \phi, \Theta, \Psi)$. The prior for each item i in group k can be defined as:

$$\Theta_{k,ii}^{-1} \sim \gamma(1, .5)$$

$$\mu_{ki} \sim N(0,1)$$

$$\nu_{ki} \sim N(0,1)$$

$$\Psi_k \sim N(0,1)$$

where $\Theta_{k,ii}^{-1}$ is the residual covariance, μ_{ki} is the intercept, ν_{ki} is the factor loading and Ψ_k is the latent mean.

Next, we assigned Δ_ν, Δ_τ to denote the cross-group differences in factor loadings and item intercepts, so that the priors of their coefficients for each item i follow:

$$\beta_{\nu i} \sim N(0, \lambda_{\nu i}^2 \tau^2)$$

$$\beta_{\tau i} \sim N\left(0, \lambda_{\mu i}^2 \tau^2\right)$$

where $\lambda_i \sim C^+(0,1)$ that is commonly used in the HS prior. For the global shrinkage parameter τ, we chose a $C^+(0,1)$ that was previously proved to yield a good performance in Bayesian variable selections (Carvalho et al., 2009; Gelman, 2006).

For the MGCFA model identification, the current study followed the reference indicator (RI) approach which had been recommended for MI testing (Rensvold & Cheung, 2008; Yoon & Millsap, 2007). Specifically, the mean and variance of the reference group were set to zero and one, respectively. Additionally, one invariant item was selected as RI and constrained to be equal between groups. we followed the approach proposed by Shi et al., (2017) to select a referent indicator (RI). Then, the model was fitted using the Stan sampler from Rstan (Stan Development Team, 2020; R Core Team, 2014) with 3000 iterations and four chains to obtain standardized posterior samples for each parameter of interest (i.e., β for all Δs). Convergence of Monte Carlo chains was assessed using the potential scale reduction

factor threshold of $\hat{R} < 1.1$ (Gelman & Rubin, 1992). All models estimated here converged, with all $\hat{R}$ values below 1.09.

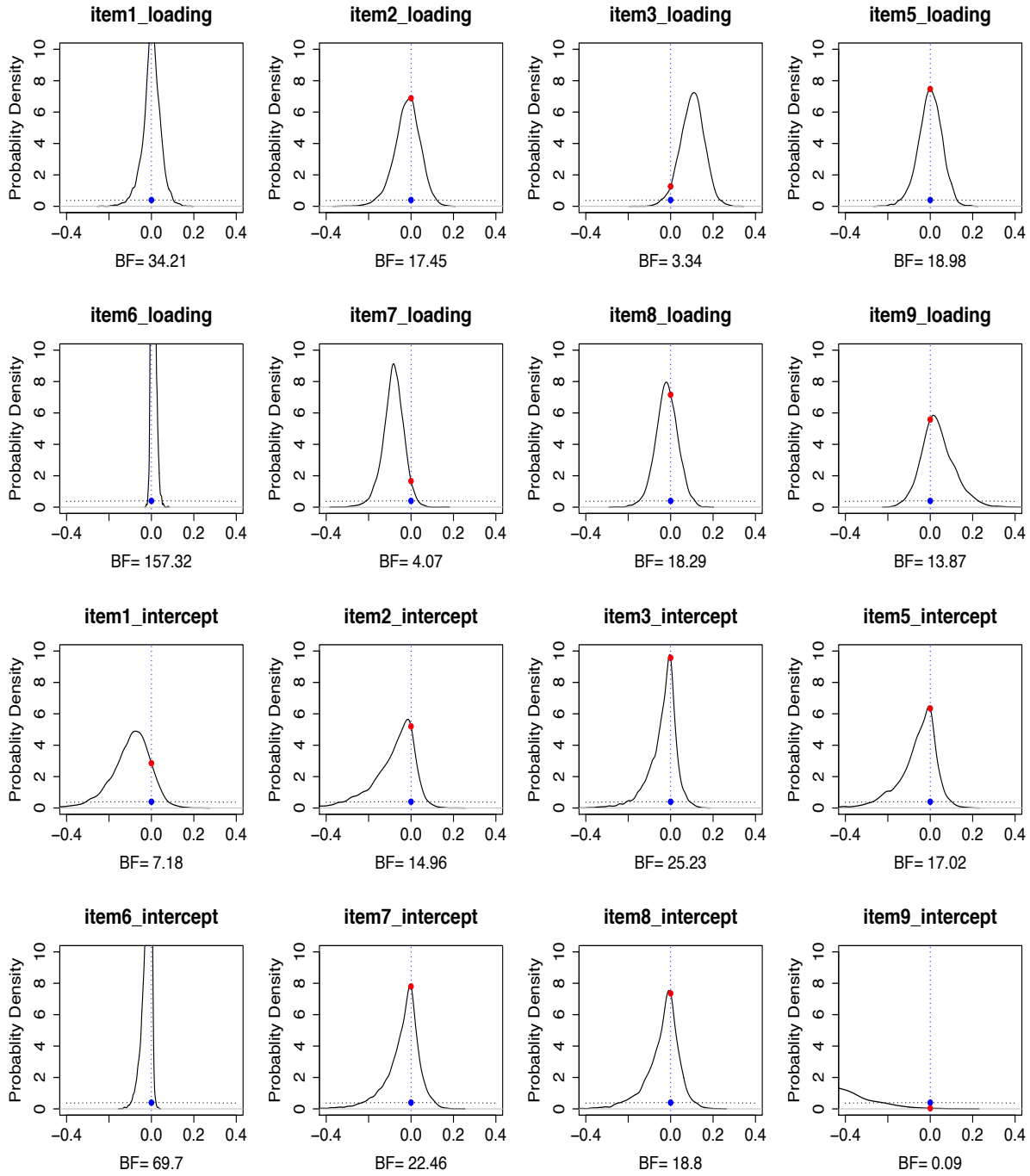
For MI assessment, the prior corresponding to the non-invariance hypothesis (H_1) was set to standard normal, and the posterior density at β was estimated using the logspline estimator. The BF_{01} was computed as the ratio of posterior and prior density at zero, for each parameter. The BF was then used to support invariance (BF_{01} > cutoffs), non-invariance (BF_{01} < 1/cutoffs), or neither due to the uncertainty in the data (Jeffreys, 1961).

4.2.1 Detecting non-invariant items Over Time

First, we evaluated the MI on CES-D 9 across the second and third waves of HRS study. From the total sample of 16,781 participants, we randomly selected a subsample of 300 participants who were repeatedly measured on both, the second and the third wave of the study. Using Shi's RI approach, we identified two items (i.e., item 3 and item 4) that qualified as RI's. Thus, we conducted MI testing twice using both items as RI and the results appeared to be identical. Here, we only reported the result when item 4 was used as RI. For clarity, we presented the posterior distributions for each parameter of interest (i.e., β for all Δs) and the corresponding BF_{01} for each loading and intercept in Figure 4. 1. A BF_{01} of 3 was used as the decision-making criterion. That is, for the decision of invariance, a BF_{01} equal or larger than 3 suggests that the probability of invariance is at least three times more likely than non-invariance. For the decision of non-invariance, a BF_{01} equal or less than 1/3 suggests that the probability of non-invariance is at least three times more likely than invariance. A BF_{01} with any values between 1/3 to 3 suggests an uncertain result. Following this rule, the results showed that factor loading invariance was established for all items. Intercept invariance was established for all items except item 9 which had a Bayes factor of 0.09. That is, item 9 was

detected as non-invariant across two measurement occasions. Therefore, a strong partial MI was established for CES-D 9 across two measurement occasions.

Figure 4. 1

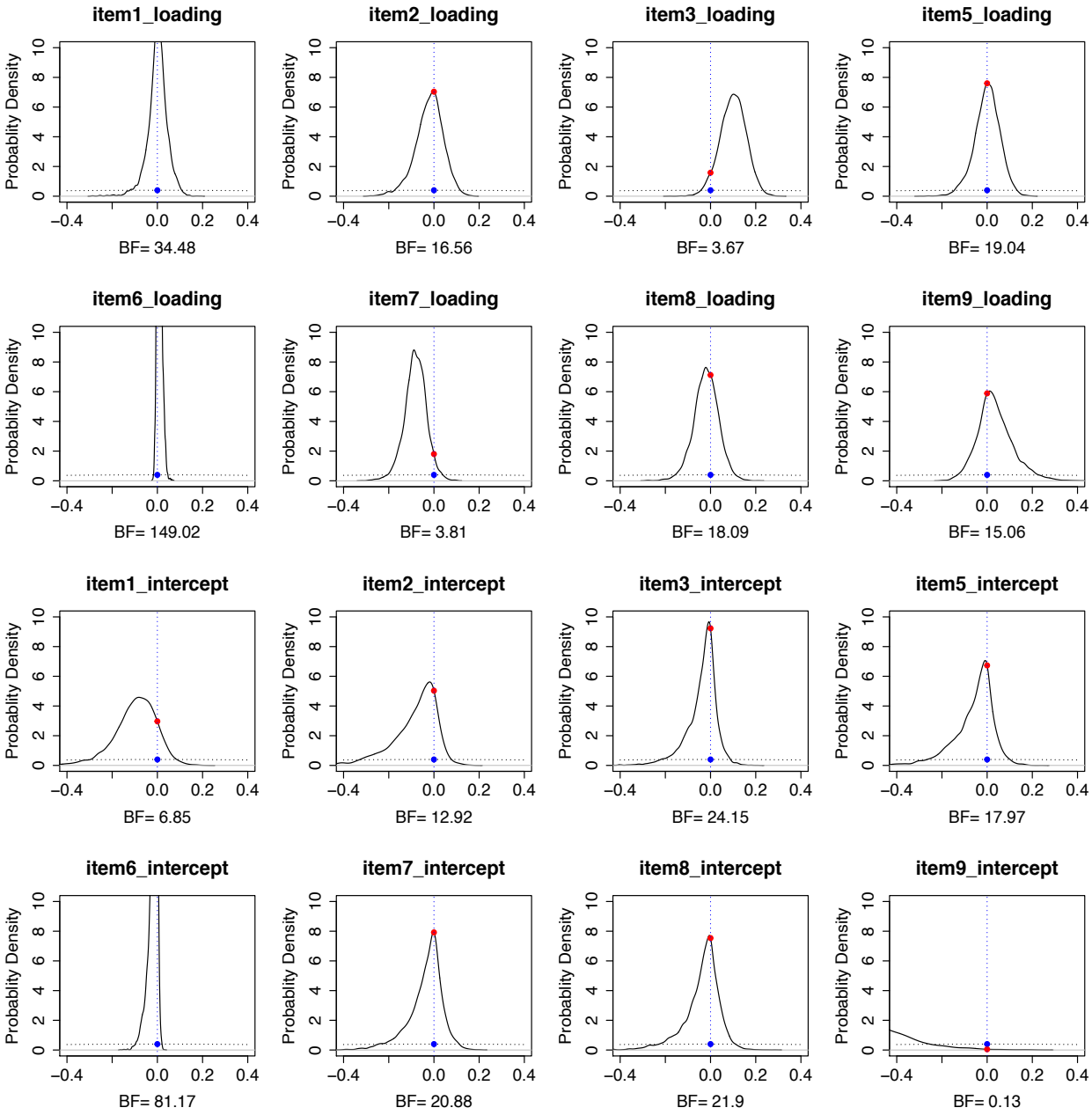


4.2.2 Detecting non-invariant items between Genders

Next, we evaluated MI on CES-D 9 between male and female participants from the second wave of HRS study. A total sample of 600 (with 50% male) participants was randomly selected. Item 4 was identified as the only RI and thus was selected for model identification. we present the posterior distributions for each parameter of interest (i.e., β for

all Δs) and the corresponding BF_{01} for each loading and intercept in Figure 4. 2. A BF_{01} of 3 was used as a decision rule here as well. The results showed that factor loading invariance was established for all items. Intercept invariance was established for all items, except item 9, which had a Bayes factor of 0.13. Again, item 9 was flagged as non-invariant between genders. Therefore, a strong partial MI was established for CES-D 9 across between male and females measured on the second wave.

Figure 4. 2



4.3 Improving the prediction of Self-deprecation to Peer Victimization

Next, we provided an empirical analysis to illustrate the use of the horseshoe (HS) prior in a partially invariant MGSEM. The data were collected from 1257 undergraduate college students (69.9% women and 30.1% men) from the California State University, Northridge. For the current application, a total of 728 participants were randomly sampled (with 50% male) for computational efficiency. Two theoretically related variables: self-

deprecation and peer victimizations were used. Self-deprecation describes the degree to which a person belittles himself/herself and his/her abilities to succeed (Owens, 1994), and was measured with the five negatively worded items from the Rosenberg Self-Esteem Scale (Rosenberg, 1979). Peer victimization was defined as repeated maltreatments (verbal, emotional, physical attacks) from one's contemporaries which was measured by a 10-item scale and an average score was used in the current analysis. (Champion & Clay, 2007). Previous research has shown that people who have negative self-views were more likely to suffer from depression (Quilty et al., 2006), which consequently made individuals more vulnerable to peer victimizations (Crick and Bigbee, 1998). Therefore, a multi-group single factor MGSEM was specified to examine the impact of latent construct self-deprecation on the manifest variable peer victimization between male and female college students.

For the current analysis, two steps were taken as following: 1) identifying the non-invariant items using the HS prior; 2) comparing the predictive accuracy of the HS prior MGSEM model with a freely estimated MGSEM model. First, we defined priors for all parameters in a one factor MGSEM model with two groups[3]. Bayesian models resulted in posterior distributions on parameters as functions of the data, likelihood, and priors: We first define priors for all parameters as following:

$$\Theta_{k,ii}^{-1} \sim \gamma(1,.5)$$

$$\mu_{ki} \sim N(0,1)$$

$$\nu_{ki} \sim N(0,1)$$

[3] To be noted, this was done for both HS model and free model.

$$\Psi_k \sim N(0,1)$$

$$\alpha_{0k} \sim N(0,1)$$

$$\alpha_{1k} \sim N(0,1)$$

$$\zeta_k \sim \gamma(1,.5)$$

where i refers to the number of items in the measurement model, k refers to the group membership. Hence, $\Theta_{k,ii}^{-1}$ is the residual covariance among items, μ_{ki} is the item intercept, ν_{ki} is the item loading and Ψ_k is the latent mean. For the s α_{0k} is the constant, $\alpha_{1\mathrm{k}}$ is the regression coefficient and ζ_k is the residual variance.

Next, we assigned $\Delta_\nu, \Delta_\tau$to denote the cross-group differences in factor loadings and item intercepts in the HS model, so that the priors of their coefficients for each item i follow:

$$\beta_{\nu i} \sim N(0, \lambda_{\nu i}^2 \tau^2)$$

$$\beta_{\tau i} \sim N\left(0, \lambda_{\mu i}^2 \tau^2\right)$$

where $\lambda_{\mathrm{i}} \sim C^+(0,1)$ that is commonly used in the HS prior. For the global shrinkage parameter τ, we chose a $C^+(0,1)$ that was previously shown to yield a good performance in Bayesian variable selection (Carvalho et al., 2009; Gelman, 2006). Therefore, the resulting posterior distributions for parameters is proportional to the likelihood, and the priors: $p(\Lambda, \phi, \Theta, , \alpha_0, \alpha_1, \zeta \mid x) \propto p(x \mid \Lambda, \phi, \Theta, , \alpha_0, \alpha_1, \zeta)\, p(\Lambda, \phi, \Theta, , \alpha_0, \alpha_1, \zeta)$.

For the MGSEM model identification (i.e., the HS model and the free model), the procedure followed the reference indicator (RI) approach which had been recommended for MI testing (Rensvold & Cheung, 2008; Yoon & Millsap, 2007). Specifically, the mean and variance of the reference group were set to zero and one, respectively. Additionally, one invariant item was selected as RI and constrained to be equal between groups. We followed the approach proposed by Shi et al., (2017) for RI selection. Then, the model was fitted using the Stan sampler from Rstan (Stan Development Team, 2020; R Core Team, 2014) with 3000 iterations and four chains to obtain standardized posterior samples for each parameter of interest (i.e., β for all Δs). Convergence of Monte Carlo chains was assessed using the potential scale reduction factor threshold of $\hat{R} < 1.1$ (Gelman & Rubin, 1992). All models estimated here converged, with all $\hat{R}$ values below 1.09.

For detecting the non-invariant items, we followed the same approach as in the previous example. That is, the prior corresponding to the non-invariance hypothesis (H_1) was set to standard normal, and the posterior density at β was estimated using the logspline estimator. The BF_{01} was computed as the ratio of posterior and prior density at zero, for each parameter. The BF was then used to support invariance (BF_{01} > cutoffs), non-invariance (BF_{01} < 1/cutoffs), or neither due to the uncertainty in the data (Jeffreys, 1961).

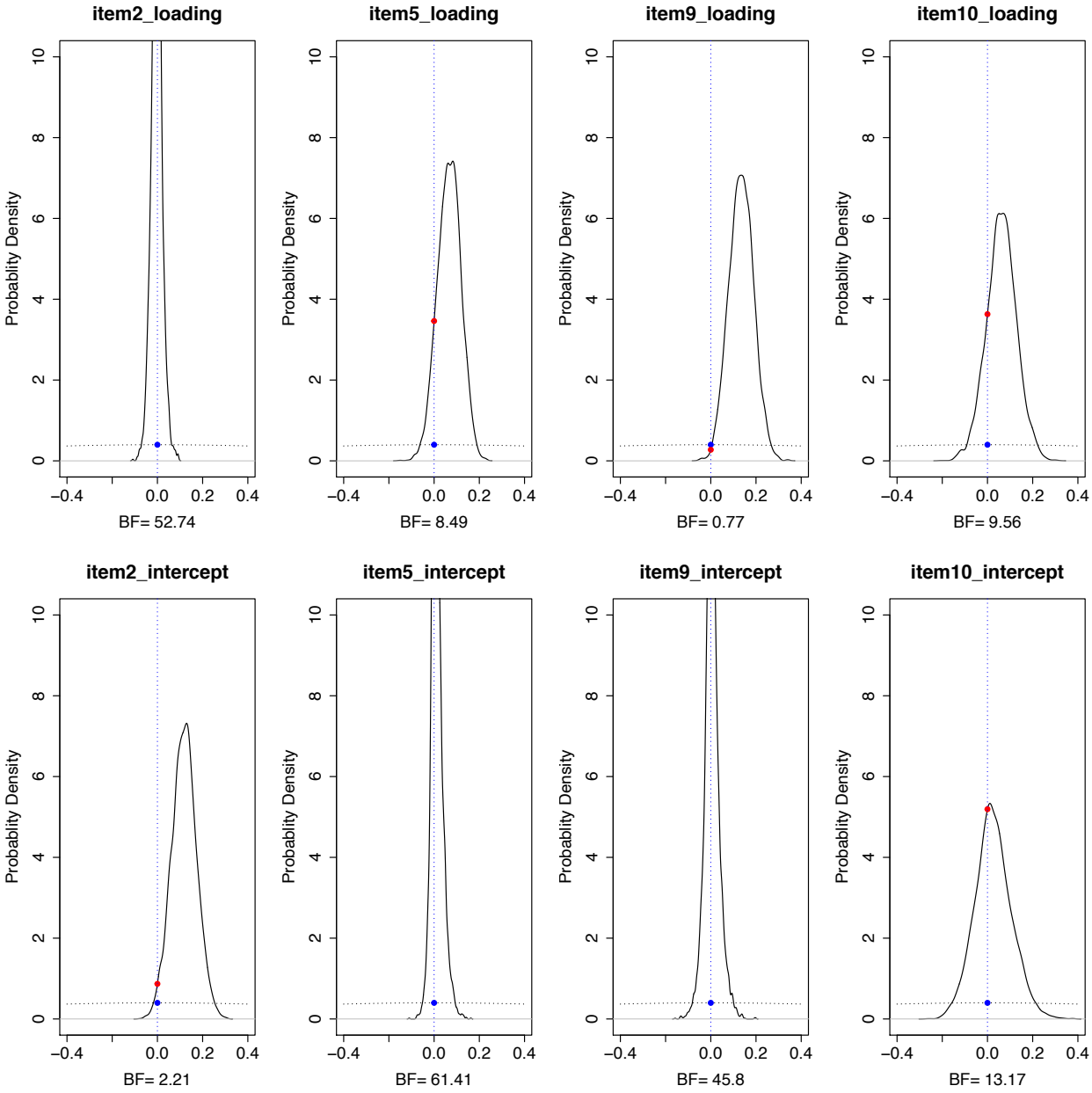
To compare the predictive performance of the HS model with the free model, two evaluation approaches were employed. First, two models were assessed on their in-sample predictions. Specifically, the Bayesian posterior predictive distributions for each model were obtained and plotted against the original data. The closer the posterior distribution get to the original data, the better predictive performance the model has. Second, two models were assessed on their out-of-sample predictions. Specifically, data were split into a training set

(about 60% of the original sample), which was used to fit models, and a testing set (the remaining 40% of the original sample), which was used to compute $\widehat{elpd_{psis-loo}}$ which approximates $\widehat{elpd_{loo}}$ for computational efficiency (Vehtari, Simpson, Gelman, Yao & Gabry, 2015). Then, we obtained the difference of $\widehat{elpd_{psis-loo}}$ between two models: $\widehat{\Delta elpd_{psis-loo-AB}}$ and its corresponding standard error: $se(\widehat{\Delta elpd_{psis-loo-AB}})$. A ratio between the difference in $\widehat{elpd_{psis-loo}}$ and the standard error of the difference: $\frac{\widehat{\Delta elpd_{psis-loo-AB}}}{se(\widehat{\Delta elpd_{psis-loo-AB}})}$ was used to select the model with better performance. A rule of thumb $\frac{\widehat{\Delta elpd_{psis-loo-AB}}}{se(\widehat{\Delta elpd_{psis-loo-AB}})} > 2$ indicated that one model performed significantly better than the other in terms of out-of-sample prediction (Bürkner, & Vuorre, 2018; Vehtari, Gelman, & Gabry, 2017).

4.3.1 Non-invariant Items

First, the non-invariance of each parameter on self-deprecation between genders was evaluated. Using Shi's RI approach, item 1 was identified and used as RI in the subsequent analysis. For clarity, we presented the posterior distributions for each parameter of interest (i.e., β for all Δs) and the corresponding BF_{01} for each loading and intercept in Figure 4. 3. A BF_{01} of 3 was used as the decision-making criteria. That is, for the decision of invariance, a BF_{01} equal or larger than 3 suggests that the probability of invariance is at least three times more likely than non-invariance. For the decision of non-invariance, a BF_{01} equal or less than 1/3 suggests that the probability of non-invariance is at least three times more likely than invariance. A BF_{01} with any values between 1/3 to 3 suggests an uncertain result. Following this rule, our results showed that factor loading invariance was established for all items except item 9 which had an impartial BF_{01} value of 0.77. Intercept invariances were established for all items except item 2 which has an impartial BF_{01} value of 2.21.

Figure 4. 3



4.3.2 Prediction between genders

Next, since the self-deprecation was found to be partially invariant between male and female college students, two model fitting strategies, 1) the HS prior model and 2) the freely estimated model, were taken to accommodate this situation using Bayesian statistics. The

plots of Bayesian posterior predictive distributions4 of both the HS model and the free model were presented for male and female separately against the actual distributions, where the original data were plotted by the thick blue line and the posterior draws were plotted by the light blue line. As shown in Figure 4. 4 and Figure 4. 5 for male college students, there was no visual difference of in-sample predictions between two models. Both models did equally well in terms of in-sample predictions given that most of the original data were covered. However, compared to the distribution of the original data, the distribution of posterior draws was slight wider and flatter which indicated an imprecision.

[4] 200 samples out of 6000 were randomly selected for plotting

Figure 4. 4

Bayesian posterior predictive distribution of the HS prior model for males

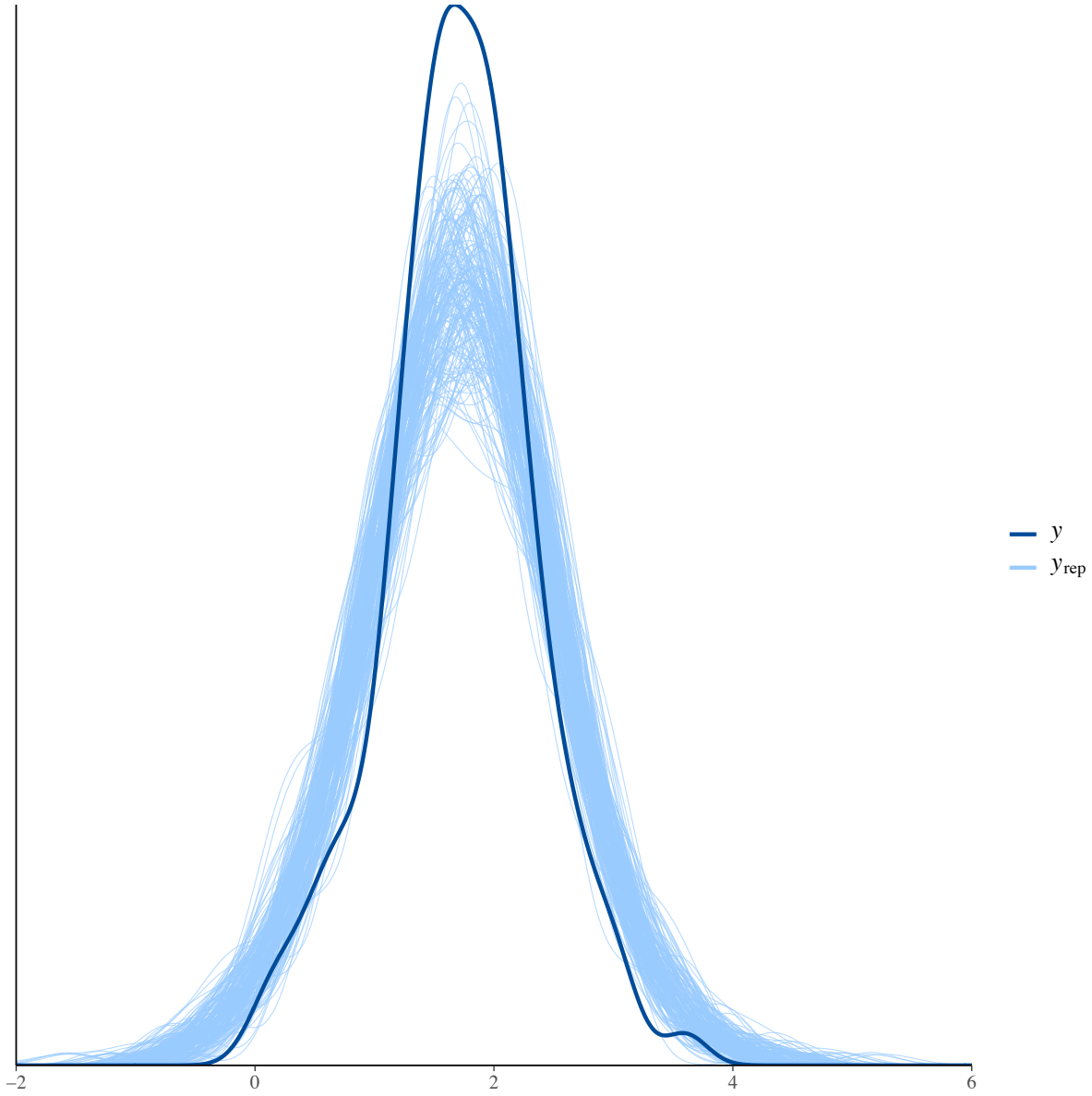
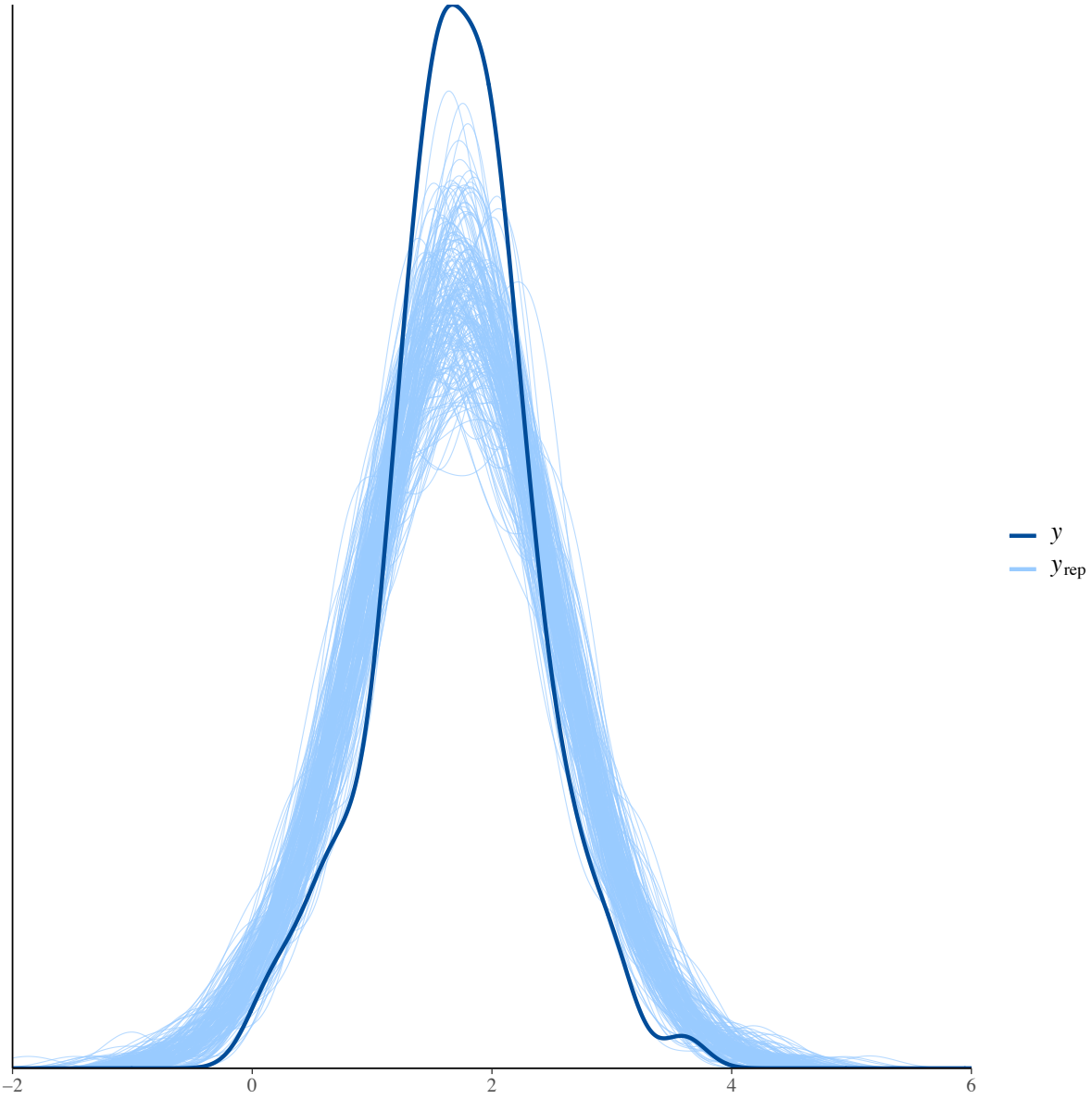


Figure 4. 5

Bayesian posterior predictive distribution of the free model for males



Similarly, for female college students, there was no visual difference in terms of in-sample prediction between two models (see Figure 4. 6, Figure 4. 7). However, comparing to the posterior prediction of male college students, the posterior prediction of female college students was less precise compared to the original data. As shown in Figure 4. 6 and Figure 4. 7, there were more posterior predictive draws fall on both tails of the distribution and the mass in the middle part was missed.

Figure 4. 6

Bayesian posterior predictive distribution of the HS prior model for females

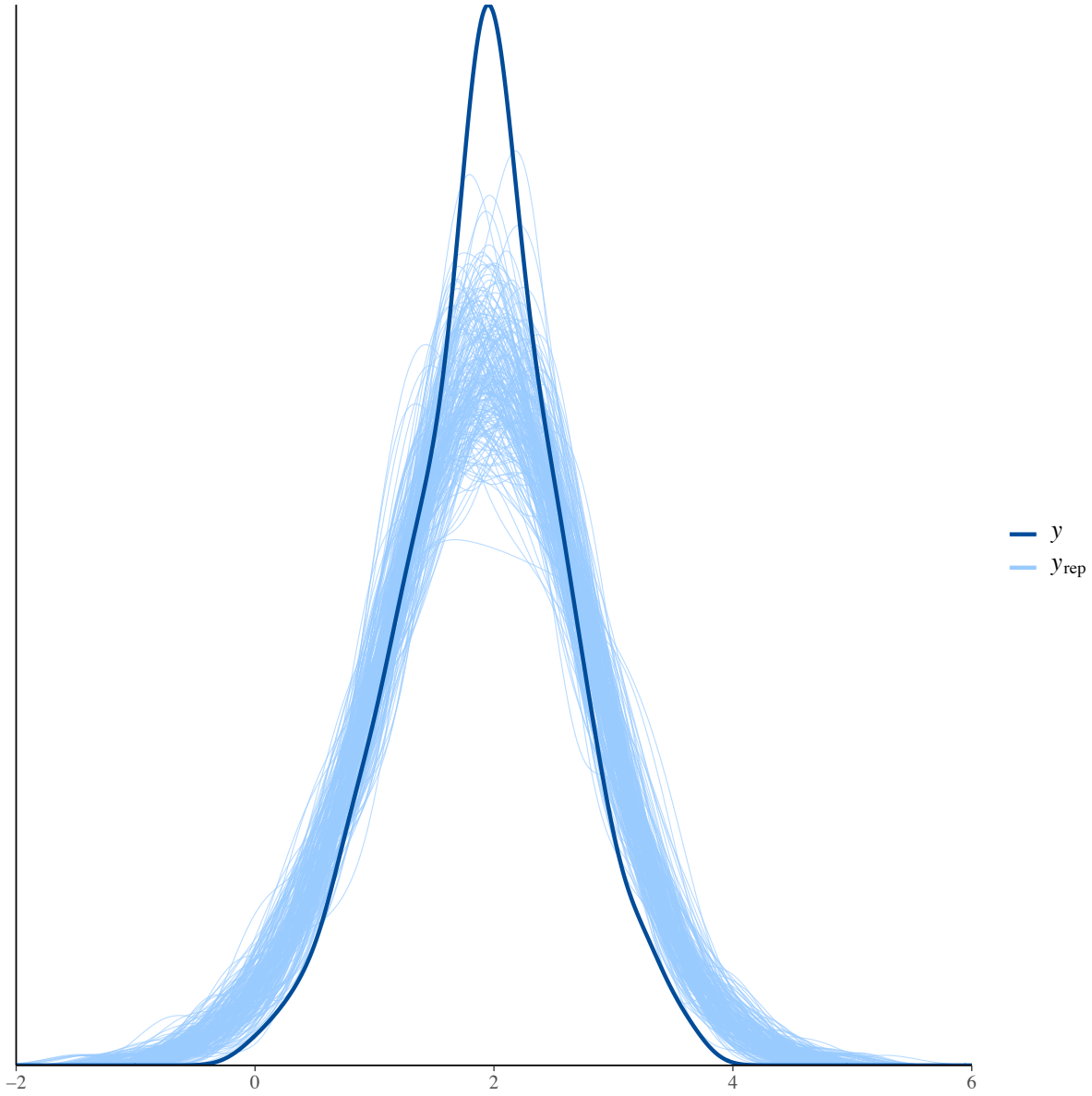
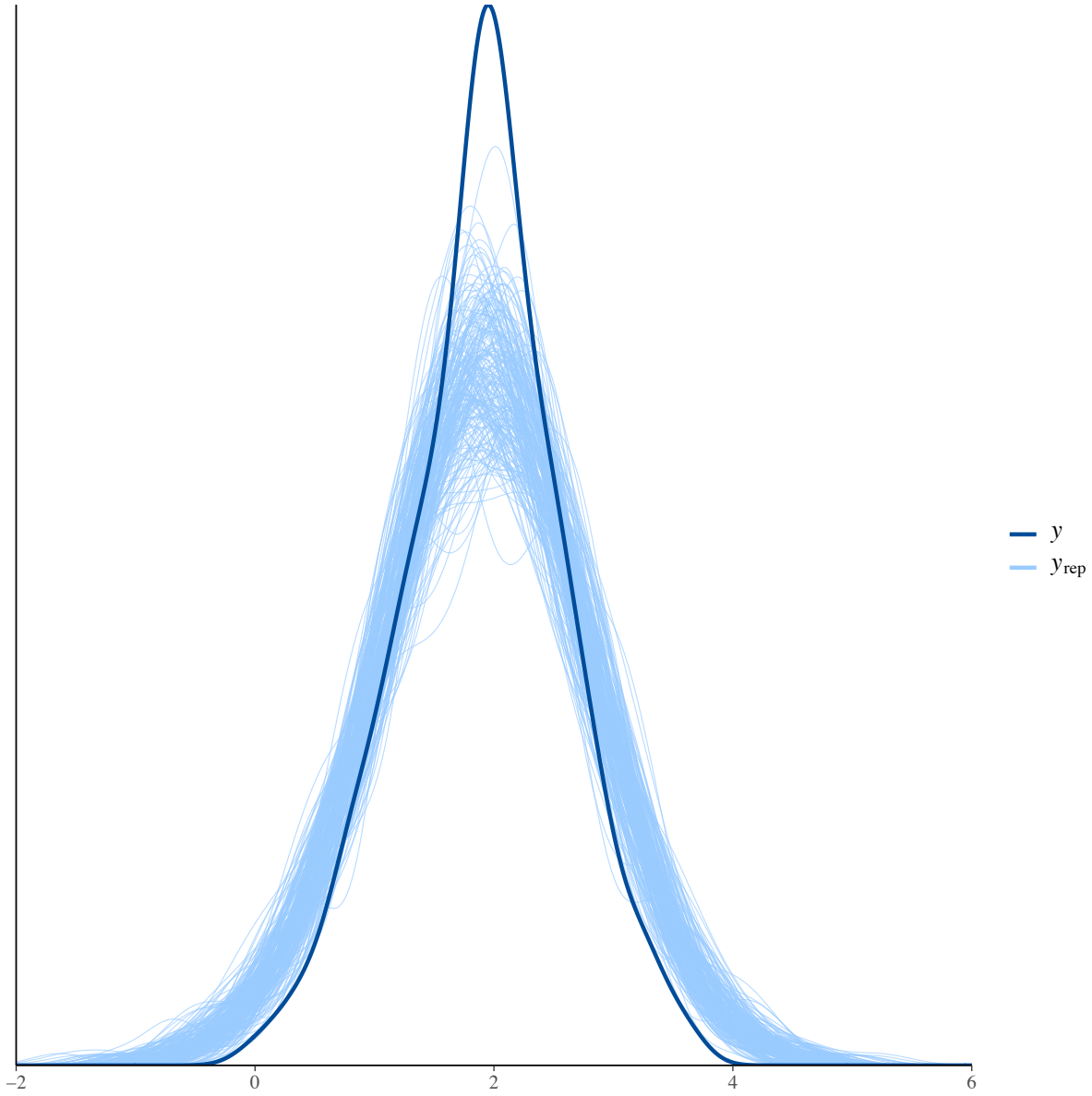


Figure 4. 7

Bayesian posterior predictive distribution of the free model for females



Next, a difference of $\widehat{elpd_{psis-loo}}$ between two models were approximated using the testing data set. The result showed that the $\widehat{elpd_{psis-loo}}$ of the HS model was significantly higher than the freely estimated model in terms of out-of-sample predictive accuracy, with a $\widehat{\Delta elpd_{psis-loo-AB}}$ of -145 (*se* = 65.5) and $\frac{\Delta elpd_{psis-loo-AB}}{se(\widehat{\Delta elpd_{psis-loo-AB}})} = 2.22 > 2$ according to the rule of thumb ((Bürkner, & Vuorre, 2018; Vehtari, Gelman, & Gabry, 2017)). Therefore, the

HS model did significantly better than the free model in terms of out-of-sample prediction on peer victimization among college students.

4.4 Discussion

In this chapter, we provided two empirical examples where we showed how to solve the issues of MI using the HS prior approach. In the first example, two MI assessments were conducted on the DER-9 item scale. As a result, item 9 was identified as non-invariant with BFs below 1/3 in both assessments. This result suggested that item 9 was problematic within the given sample since individuals from different groups or at different measurement occasions responded differently on this item. Therefore, researchers could either get this item deleted, or investigate the content of this item and make some revisions accordingly. For those researchers who wish to proceed with the analysis without deleting any items, the second example would be quite enlightening. In the second example, we showed how to use the HS prior to remedy a partially invariant MGCFA model that was for prediction. First, a MGCFA model measuring self-deprecation was tested for invariance and was found to only hold for partial invariance. Next, a MGSEM model was specified where a manifest variable "peer victimization" was regressed on "self-deprecation". To accommodate partial invariance, all the cross-group constraints were substituted by the HS prior, which implying that given the data, the amount of non-invariance would be allowed if it contributes significantly, otherwise it would be shrunken towards zero. For comparison, a freely estimated MGSEM model where no group constraint was imposed was also specified. Overall, the HS model performed significantly better than the free model in terms of out-of-sample prediction.

Further implications and limitations of this study are addressed in the next and final chapter.

5. Chapter 5

General Discussion

5.1 Overview of Findings

In this set of studies, we sought to answer questions related to measurement invariance (MI) under the structural equation modelling (SEM) framework. Our goal was to address the following questions related to MI assessment: (a) Can we treat the detection of non-invariant items as a variable selection problem by introducing a set of variables representing cross-group differences?; (b) Can we assess the effectiveness of cross-group differences via a Bayesian variable selection approach using the horseshoe (HS) prior? ; (c) Can we gather the evidence of measurement invariance/non-invariance using Bayes factors? In addition, we also attempted to solve the following questions concerning partially invariant models: (a) When estimating partially invariant models, what is the major risk when only a single model fitting strategy is considered? (b) Can we employ Bayesian-model-averaging (BMA) to take advantage of all potential models instead of selecting a single model to improve model predictions? (c) Can we use the HS prior to substitute BMA in estimating partially invariant models for computational efficiency? we investigated each of these questions in two separate studies. In the third study, we provided two empirical analyses on two set of multigroup measurement data.

5.1.1 Study 1

The goal of Study 1 (Chapter 2) was to introduce a new method in assessing MI. we reframed item-level MI assessment as a variable selection problem where we defined a set of variables representing cross-group differences and placed the HS prior on this set of variables to mimic the conditions of non-invariance and invariance. Bayes factors (BF) were used as a decisions tool where the empirical evidence of invariance and non-invariance on each

parameter were gathered simultaneously. To test this method, we simulated data based on a single factor MGCFA model which varied in sample size, item reliability, magnitude of difference on factor loadings and intercepts and scale length which resulted in 72 unique conditions. Each of these data conditions corresponded to different non-invariant situations such as a small sample size with large non-invariances and a low item-reliability, or a large sample size with small non-invariances and a medium item-reliability. The purpose was to investigate the performance of the HS prior in detecting non-invariant items under different non-invariant situations. Then we fitted MGCFA models with the HS prior to the simulated data and evaluated the accuracy of non-invariance detection across different conditions. Five metrics were used for evaluating the performance of the HS prior approach: 1) *Certainty* which describes the proportion of all comparisons yielding certain decisions in determining invariance versus non-invariance using BF. 2) *Sensitivity* which describes the proportion of all non-invariant parameters detected as non-invariant. 3) *Specificity,* defined as the proportion of all invariant parameters detected as invariant. 4) *Positive predictive value* (PPV) which is defined as the proportion of all parameters detected as non-invariant that are truly non-invariant. 5) *Negative* predictive value (NPV) that is defined as the proportion of all parameters detected as invariant that are truly invariant.

Results from Study 1 suggested that the HS prior approach performed well in assessing item-level invariance. Overall, sample size, item reliability, and the magnitude of non-invariance showed notable impacts on MI testing, while scale lengths and the pattern of non-invariance (i.e., whether items hold for loading non-invariant or intercept non-invariant or both) did not. First, high Certainty rates (i.e., no less than 85%) indicated that the decisions of invariance versus non-invariance could be made for certain under all conditions. A low item-reliability along with small non-invariance led to a large portion of undecided results.

Next, this method showed an excellent performance in correctly detecting invariant items across all conditions which was signaled by perfect specificity rates and high PPV. It should be noted that specificity and PPV were only affected by sample size and item reliability but not the magnitude of non-invariance. This suggested that a truly invariant parameter was almost always correctly classified as invariant regardless of sample size and item reliability.

In contrast to detecting invariance, detecting non-invariance seems to be more difficult via the HS prior approach. A large variation in sensitivity among different simulation conditions suggested that the identification of non-invariance heavily depended on sample size, item reliability and the magnitude of non-invariance. In the worst case scenario (i.e., small sample size, low item reliability, small non-invariance), less than 10% of non-invariances detections were successful. As sample size and item reliability increased, the detection of small non-invariance became easier. Large non-invariances were more readily detectable despite of smaller sample sizes and lower item reliability. Although detecting non-invariance was overall challenging, high NPV implied that once a parameter was identified as non-invariant, the identification was generally accurate. This was a similar finding as for the detection of invariance.

To recap, we attempted to conduct item-level MI assessment via a Bayesian variable selection method using the HS prior and collect the evidence of invariance on each parameter using BF in this study. The results from Study 1 suggested the identification of invariance was mostly successful and accurate when BF yielded certain decisions. On the other hand, the detection of non-invariance frequently failed under some disadvantageous conditions (e.g., small sample size) in the sense that the evidence of non-invariance was insufficient. This failure could be explained by a small effect size problem. That is, small effects (i.e., small non-invariance) were hard to manifest with a small sample size and a highly noisy

environment (i.e., item reliability). In an empirical setting, these findings implied that small differences in parameters might be hard to capture and had small impact on the overall estimation when sample size was small and item reliability was low, so that might be reasonably ignored.

5.1.2 Study 2

In Study 2 (Chapter 3), a BMA approach was introduced as an alternative modelling strategy other than relying on a single model when the MI assumption cannot be fulfilled. Conventionally, researchers will attempt to select a best model fitting strategy among a set of candidate models when MGCFA models only hold for partial invariance. The drawback of this model selection approach is that it overlooks the uncertainties among all potential modelling strategies. Consequently, the statistical inference relying on a single model solution might result in an overly confident conclusion about its parameter estimates. To overcome this issue, we employed Bayesian-model-averaging (BMA) technique which was accomplished by placing the HS prior on a set of parameters representing non-invariance in each parameter. In theory, by averaging over the entire model space, the results should yield best predictions compared to any single best model solution. To test this method, a population model was simulated following a multi-group SEM where a manifest variable was regressed on a single latent factor model. The simulation condition was identical to Study 1 where sample size, scale length, the magnitude of non-invariance, pattern of non-invariance and item reliability were manipulated to yield a total of 72 conditions. we used the approximated expected logarithm pointwise predictive density ($\widehat{elpd_{psis-loo}}$) as a measure of the models' out-of-sample prediction. Three common modelling strategies: a freely estimated model, a partially constrained model and a fully constrained model were used for comparison in terms of models' predictive ability. The purpose was to see if the HS prior model indeed

led to a better prediction than the other models. Then we fitted each model to the simulated data and evaluated their out-of-sample prediction via $\widehat{elpd_{psis-loo}}$ across different conditions. One metric was used for model comparisons. Absolute best rate was defined which describes the proportion of each model being significantly better than other models according to the rule of thumb: $\frac{\Delta\widehat{elpd_{psis-loo-AB}}}{se(\Delta\widehat{elpd_{psis-loo-AB}})} > 2$.

Findings from Study 2 suggested that the HS prior model led a better out-of-sample prediction comparing with other models in most simulation conditions. In general, the HS prior model has a significantly higher $\widehat{elpd_{psis-loo}}$ than the other models for most of the simulation conditions. The magnitude of non-invariance and item reliability all showed impacts on the *absolute best* rate of the HS prior model. In most simulation conditions, a large sample size, a high item reliability, and a small magnitude of non-invariance led to a higher *absolute best* rate of the HS prior model. In contrast, when the sample size was small, the item reliability was medium or low, and the magnitude of non-invariance was large, the performances of the HS prior model and the partially constrained model were almost indistinguishable. Similarly, the HS prior model and the partially constrained model performed equally well when the non-invariance on item intercepts was large, and item-reliability was low. Also, no significant difference was observed in the *absolute best* rate of the HS prior model between the 6-item scale and the 9-item scale.

In Study 2, we tried to tackle the issues of predictive performance with partially invariant models. When MGCFA models only hold for partial invariance, the conventional approach tends to select a best fitting model which provides good estimates for current data yet may lead to poor predictions for the future data. As an alternative, we adopted the idea of the BMA technique which takes the information from all possible models via averaging the

entire model space. Yet, instead of implementing the actual BMA, the HS prior approach was chosen to simplify the estimation procedure. Specifically, instead of having to estimate all possible models, the HS prior approach weighs individual model parameters and results in an equivalent solution as the one obtained with BMA. The results of Study 2 confirmed the usefulness of the HS prior approach in improving the predictive ability of partially invariant models. However, it should be noted that under certain conditions, especially when item reliability was low and the magnitude of non-invariance on intercepts was large, the HS prior model sometimes failed to outperform the partially constrained model. That is, both modelling strategies performed equally well in terms of predictions. This is likely due to the higher flexibility of the HS prior in accommodating non-invariance based on their magnitudes, while the partially constrained model only made dichotomous decisions. So, when the non-invariance was small and negligible, relaxing constraints might be unnecessary and shrinking the non-invariance could lead a better prediction. Once the magnitude of non-invariance increased, relaxing constraints might produce similar results as the HS prior. Also, it may be also due to that the local shrinkage and global shrinkage parameters used in the current HS prior can better handle the simulated values in factor loadings but not intercepts. Although under certain conditions the HS prior model performed equivalently to the freely estimated model, it should be noted that the true non-invariant parameters were only known in simulation study. Fitting a partially constrained model requires the correct identification of non-invariant parameters which can never be guaranteed in empirical studies. By contrast, the HS prior model can identify the non-invariant parameters and automatically adjust the constrained values accordingly. Thus, the HS prior should be a preferable way in using partially invariant models for predictions.

5.1.3 Empirical Example

Finally, in Study 3 (chapter 4), we performed two empirical analyses to illustrate 1) assessing item-level invariance via the HS prior, 2) improving the prediction of partially invariant models with the HS prior. To accomplish the first goal, we conducted two MI assessments, where DERS-9 scale was assessed on MI between genders and between two measurement occasions. To achieve the second goal, we first confirmed the partially invariant status of the "self-deprecation" scale between male and female college students. Then, the HS prior model and the freely estimated model were fit to the partially invariant data where a theoretically related variable "peer victimization" was regressed on "self-deprecation" between genders. The predictive performance of both models was then compared on in-sample predictions via Bayesian posterior predictive distributions. To compare out-of-sample predictions between two models, about 60% of data was used to fit models, and 40% of the data was used to compute $\widehat{elpd_{psis-loo}}$.

Findings from the first part of this study suggested that the HS prior approach functioned efficiently and precisely in identifying non-invariant parameters. In both MI tests, item 9 was flagged as problematic item with a non-invariant intercept and so that DERS-9 was only held for partial invariance. In the context of research, several options are available after the source of non-invariance was identified. For psychometricians whose primary goals are identifying problematic items and maintaining the quality of scales, they could either choose to delete or edit problematic items. For clinical practitioners, whose primary goal is using scales for candidates' selection or predicting individuals psychological well-beings, they could either select a model fitting strategy to accommodate partially invariant scales or take the BMA approach which was implemented in the second part of this study. Specifically, the second part of Study 3 explored the possibility of utilizing the HS prior to

improve the prediction of partially invariant models. The finding of Study 3 was in line with the simulation result from Study 2, as the HS prior model generally did better than the freely estimated model. Although both models were almost indistinguishable in terms of in-sample predictions, the HS prior model indeed yielded significantly higher predictive probability than the freely estimated model for out-of-sample predictions.

5.2 Limitations & Future Directions

There are a number of limitations in this set of studies that should be addressed and expanded in future research. One limitation is that the choice of BF cutoff is arbitrary, just like the choice of an alpha level (or p value) in null hypothesis significant testing (Raftery, 1995; Wagemaker, 2007). Although they differ in practical implication and statistical interpretation, both p value and BF reflect the level of researchers' confidence in making decisions. For instance, if we are confident in making a decision when one hypothesis is 3 times more likely than the competing hypothesis, choosing a relatively low Bayes factors cutoff (i.e., BF = 3) will be acceptable. If we think even stronger evidence is needed when making decisions (e.g., medical research, clinical trials), choosing a larger Bayes factors cutoff that exceeds 20 may be more reasonable (cf. Raftery, 1995). This is consistent with our simulation results, where the increasing threshold of Bayes factors leads to decreasing numbers of decisions that can be made. We recommend the BF of 3 as a reasonable choice for decision making criterion in assessing MI, given that choice of BF cutoff only impacts the certainty rate, but not the decision accuracy in our simulation. However, as the practical significance of statistical results may vary across research areas and from studies to studies, researchers should adjust this value accordingly.

Another limitation of the current study is that the accuracy of non-invariant/invariant item detection is conditional on the unbiased referent indicators selection. That is, referent

indicators (RI) that are imposed with equal-group constraints must be truly invariant. Otherwise, if a non-invariant item is taken as referent indicator, the subsequent MI assessment may fail and invariant items may be falsely flagged as non-invariant, or vice versa (Shi et al., 2007; Rensvold & Cheung, 1998). Several methods have been proposed under both Bayesian and frequentist frameworks in addressing RI selection issue (Cheung & Lau, 2012; Cheung & Rensvold, 1998; Kim & Yoon, 2011; Shi et al., 2007; Stark, Chernyshenko, & Drasgow, 2006). For simulation studies, the invariance states of items were known so that there was no need to search for RIs. For empirical analyses, we used the method by Shi et al., (2017) to identify potential RIs since it was the only one taking a Bayesian approach. Future studies could explore other RI searching methods under a Bayesian framework such as extending Shi et al., (2017) study by employing the HS prior.

Another limitation of the current study is associated to the prior choices of hyperparameters in the HS prior. Previous studies have discussed different prior choices for the global shrinkage hyperparameter in the HS prior (Piironen& Vehtari, 2017). The current study employed a half-Cauchy distribution with mean of zero and variance of one (i.e., $C^{+}(0,1)$), which is the most common one, as the prior for the global shrinkage hyperparameter. Researchers have argued that this default option often left significant amounts of parameters (e.g., regression coefficients) unregularized and may result in bad results such as data separation in the logistic regression when parameters are only weakly identified (Piironen&Vehtari, 2017). However, this should not be a big concern in the current study. First, the current application of the HS prior for improving the prediction of partially invariant models served primarily as a proof of concept and as a first step into exploring its use. Second, unlike some common HS prior applications where statistical models contain a large number of parameters (e.g., a regression model with a large number of predictors), the

number of target parameters (i.e., cross-group differences) in MGCFA models are relatively small. So, it is unlikely that the problem associated with this default option will lead to the under-regularization problem[5]. Finally, the method of specifying priors for the global shrinkage parameter proposed in Piironen and Vehtari (2017) does not fit into the framework of MI studies where we cannot know the number of non-invariant parameters beforehand. Because it requires that researchers have some ideas about the number of effective parameters (Piironen & Vehtari, 2017), which can be translated as the number of non-invariant parameters in MI studies.

Another limitation is specific to the simulation design. In the current study, the simulated data were normally distributed with no missing values, equal sample size among groups, and the model structure was simple (i.e., only contained one latent factor without item cross-loadings or correlated residual variances, exogenous was observed variable), and MI assessments were only conducted between two groups. For greater generalizability, future studies should investigate the performance of the HS prior in different simulation settings, such as when there are more than two groups, the data contain missing values, sample sizes are unequal across groups, or the model structure is more complex (e.g., item cross-loadings, correlated residual variance). In addition, since our study is not the only one using Bayesian regularization methods in MI assessment (see Chen, Bauer, Belzak, & Brandt, 2021; Liang, & Luo, 2019; Muthén et al, 2013, 2017; Shi, Song, DiStefano, Maydeu-Olivares, McDaniel, & Jiang, 2019), future studies could compare the HS prior approach with other regularization methods such as the Spike-and-Slab prior approach (Chen et al., 2021), the small variance approach (Muthén et al, 2013, 2017; Shi et al., 2017), and the frequentist Lasso method

[5] We conducted several simulations with the different values of the global shrinkage parameter but did not find any difference.

(Bauer, Belzak, & Cole, 2019) and discuss some pros and cons in taking each approach in detecting non-invariant items.

Finally, in the empirical analyses, we only compared the predictive performance of the HS prior approach with the freely estimated model. Although the result of Study 2 showed that partially constrained models performed as well as the HS prior models in terms of predictions under some simulation conditions, we do believe that this can only happen in simulations where the state of invariance is known for each item. In empirical analyses, there is no guarantee that partially constrained models are always correctly specified given the false positive and false negative in non-invariant items detections. Therefore, freely estimated models are safer choices than partially constrained models in practice, and hence is more appropriate to be used for model comparison.

5.3 Recommendations for the use of Horseshoe prior with MI assessment study

Despite the limitations of this set of studies, we advocate for the Bayesian variable selection approach as a viable method in examining the item-level MI and improving the predictive ability of partially constrained models. First, one should begin with considering the main research question they want to explore with MGCFA models. If the aim is to identify problematic items (e.g., DIF), researchers should take Study 1 as a reference. Particularly, when determining the state of non-invariance for each parameter of interest, one should choose the cutoff value of BFs based on their actual needs. For instance, when strict invariance is required for practical purpose, such as using assessments for medical research or candidate selections, a large cutoff value of BFs (e.g., BFs > 10 for invariance) should be used where the decision of invariance is made upon a solid amount of empirical evidence. On the other hand, for some fields where the requirement of invariance can be relaxed a little bit (e.g., a group of school psychologists are trying to assess the level of math anxiety among

sixth graders), one can choose the default cutoff value (i.e., BFs > 3). When there is a fair number of items and none can be determined as either invariant or non-invariant (e.g., 3 > BFs > 1/3), one should contemplate if this is caused by a lack of statistical power or a low item-reliability, or both. As shown in Study 1, sample size and item-reliability are two key factors impacting the certainty rate. As sample size increases, one can expect more conclusive and accurate results. Thus, when a considerable number of items are found as undetermined, one can start on examining item-reliability and may consider increasing the sample size if it is possible. In terms of using MGCFA models for predictions only, one could simply skip the non-invariance detection and directly fit a MGSEM model following the procedures in Study 2 and Study 3. To check the quality of prediction, one way is to compare the out-of-sample prediction of the HS model with a freely estimated model using $\widehat{elpd_{psis-loo}}$ (see Study 3). Though, frankly speaking, we do not recommend this approach. Using the BMA approach to fit a partially invariant model is like putting a middle ground between the theory (i.e., strict MI) and the reality (i.e., data). On one hand, it permits some flexibilities (i.e., random noises) on the cross-group difference. On the other hand, it takes care of the non-invariance between groups. Therefore, if one could be satisfied with a freely estimated model, then why even bother testing for MI in the first place? As alternatives, we can take the in-sample prediction approach by plotting the posterior distribution of each group separately against the original data. If the posterior distribution seems way too off compared to the original data, one can go back to check the model to see if there is any misspecification, or there are too many parameters exhibiting non-invariance.

5.4 Conclusion

Psychological research often concerns the score or trait comparisons from individuals in heterogeneous conditions such as different ethnicity groups, taking evaluations at different

locations, or being assessed across time. To obtain valid scientific inferences from such comparisons, researchers need to ensure the consistency of those scores. Although the same measurement tends to be used under such a circumstance, there is no guarantee that the mean differences in the observed scores are purely driven by individuals' different states on the constructs. The desirable situation should be that the mean differences in scores truly/solely reflect the individual differences in the same psychological construct. The term measurement invariance (MI) is used to refer to this situation. The violation of MI could induce measurement bias, so that the observed trait difference may be confounded with the difference of measurement construct. Therefore, the MI testing has been developed to examine if the same trait remains unchanged regardless of the different conditions (e.g., Jung & Yoon, 2016; Mellenbergh, 1989). The establishment of MI is considered as an important premise for social and behavior research since it enables the valid cross-group comparisons. Difficulties arise, however, the establishment of full MI is often found unrealistic in practice. One reason may be that the majority of traditional MI testing methods (e.g., constrained model comparisons) tend to be overly restrictive (i.e., allowing zero difference). The increasing model constraints may result in a high model rejection rate even if the group difference is small, especially when analyzing complex models (e.g., multi-factors, correlated errors) or large scales. Up to now, MI studies have been mostly centering upon MI assessments, while less attention has been given to the potential solutions when the full MI fails.

With the failure of MI, one common practice is to locate the exact source causing non-invariance and fit a partially invariant model. Nevertheless, neither identifying non-invariant parameters, nor selecting an appropriate model fitting strategy is an easy task. And it should be noted that whether the selected model fitting strategy will yield good results

(e.g., accurate estimates, better predictions) is conditional on the correct identification of non-invariant parameters. In other words, any incorrect invariance detection could bias the results of partially invariant models. Previous MI studies mostly focused on the model level assessment which does not provide an exact location of non-invariance and potentially inflates Type I and II errors (e.g., constrained model comparison). There are only a handful of studies that have explored the direct way of detecting non-invariant parameters (i.e., item level assessment), which has not yet been widely applied. Also, there is a fundamental issue with respect to selecting one single model fitting strategy for partially invariant models. That is, the uncertainty in model selection process is largely ignored and the final selected model is believed to reflect the true data generating process. Whether a single model solution obtained from the given data can produce a better forecast for unseen data or can be utilized on the future studies is under doubt. Thus, this behavior is considered "dangerous" in the sense that it neglects the complexity of human characteristics and the evolution of science.

The present dissertation aimed to address the aforementioned two important issues related to MI studies. Importantly, we investigated the potentials of a Bayesian variable selection approach in detecting non-invariant items and a BMA approach to improve the predictive abilities of partially invariant models. In this set of studies, we first attempted to solve two issues faced by the traditional MI testing using a Bayesian approach. Specially, under the Bayesian variable selection framework, we employed a Horseshoe (HS) prior approach to enable a parameter level MI testing which not only avoided the issue of inflated error rates (i.e., both Type I and Type II) due to multiple model comparisons, but also allowed a precise non-invariance detection. Additionally, we used Bayes factors (BF) in determining the state of invariance to prevent the theoretical issue caused by null hypothesis significance testing. The simulation results showed that the Bayesian variable selection

approach performed well when sample size was sufficient, and item-reliability was acceptable. Then, we seek to improve the predictive ability of partially invariant models when one or more non-invariant parameters are detected. Specifically, most studies concerning the predictive ability of partially invariant models have focused on selecting a single model fitting strategy to obtain accurate predictions, such as deleting the non-invariant items, fitting partially constrained models, using composite scores, or fitting fully constrained models. One fundamental issue of making decisions conditional on one single model is that it overlooks the uncertainties in model selection, which is especially the case for partially invariant models. In partially invariant models, the pattern and magnitude of non-invariance can be highly complex and vary case by case. Thus, fitting a single model may not be optimal for predicting the unseen data. To overcome this issue, we again used the HS approach to mimic Bayesian-Model-Averaging (BMA) where the prediction is made upon all possible models. We expected to see partially invariant models show a higher predictive probability with the HS approach when comparing with other common model fitting strategies (e.g., partially constrained models, freely estimated models). We demonstrated this via simulation studies with a partially invariant Multi-group SEM model where four model fitting strategies were compared in terms of predictive probabilities. The HS model exhibited an excellent predictive performance and outperformed other model fitting strategies under almost all simulation conditions. Finally, two sets of empirical analyses were conducted for the illustrative purpose. In the case of non-invariant item detection, the Bayesian variable selection approach picked up the source of non-invariances in one psychological measurement from two different aspects (i.e., between groups, over time). In the case of predictions with partially invariant models, the HS prior approach showed an excellent performance with both in-sample forecast and out-of-sample forecast when comparing with other model fitting strategies.

Despite all the limitations that have been discussed in the previous section, the Bayesian variable selection approach provides a new perspective for MI studies, where non-invariance detection and partially invariant model prediction are achieved simultaneously. Although under this method, the distinction between invariance and non-invariance may become blurry, and the partially invariant model may become uninterpretable. Yet, the wisdom token "all models are wrong, but some are useful" from George E. P. Box suggests that we should focus more on "is this method useful for providing a solution to the failure of MI", instead of "is this method correctly representing the state of non-invariance". After all, measurement invariance does not mean anything by itself and only becomes meaningful when it comes to contributing to social and behavior research.

References

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. *Robustness in Statistics*, 201–236. doi:10.1016/b978-0-12-438150-6.50018-2

Bürkner, P.C., & Vuorre, M. (2018). *Ordinal Regression Models in Psychology: A Tutorial.* doi:10.31234/osf.io/x8swp

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.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009, April). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics* (pp. 73-80). PMLR.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika, 97*(2), 465–480. doi:10.1093/biomet/asq017

Champion, K. M., & Clay, D. L. (2007). Individual differences in responses to provocation and frequent victimization by peers. *Child Psychiatry & Human Development, 37*, 205–220. doi:10.1007/s10578-006-0030-9

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14* (3), 464–504. doi: 10.1080/10705510701301834

Chen, S. M., Bauer, D. J., Belzak, W. M., & Brandt, H. (2021). Advantages of Spike and Slab Priors for Detecting Differential Item Functioning Relative to Other Bayesian Regularizing Priors and Frequentist Lasso. *Structural Equation Modeling: A Multidisciplinary Journal, 29*(1), 122–139. doi:10.1080/10705511.2021.1948335

Cheung, G. W., & Lau, R. S. (2011). A Direct Comparison Approach for Testing Measurement Invariance. *Organizational Research Methods, 15*(2), 167–198. doi:10.1177/1094428111421987

Cheung, G. W., & Rensvold, R. B. (1998). Cross-cultural comparisons using non-invariant

measurement items. *Applied Behavioral Science Review, 6*(1), 93–110. doi:10.1016/s1068-8595(99)80006-3

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9* (2), 233–255. doi: 10.1207/s15328007sem0902{_}5

Crick, N., Bigbee, M., & Kendall, Philip C. (1998). Relational and overt forms of peer victimization: A multi-informant approach. *Journal of Consulting and Clinical Psychology, 66*(2), 337-347.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, *42* (1), 204–223.

Ezpeleta, L., & Penelo, E. (2015). Measurement invariance of oppositional defiant disorder dimensions in 3-year-old preschoolers. *European Journal of Psychological Assessment*, *31* (1), 45–53. doi: 10.1027/1015-5759/a000205

Finch, W. H., & French, B. F. (2018). A Simulation Investigation of the Performance of Invariance Assessment Using Equivalence Testing Procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–14. doi: 10.1080/10705511.2018.1431781

French, B. F., & Finch, W. H. (2008). Multigroup confirmatory factor analysis: Location the invariant referent set. *Structural Equation Modeling*. doi: 10.1080/10705510701758349

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*(3). doi:10.1214/06-ba117a

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457-472.

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*. doi: 10.1016/j.jmp.2015.09.001

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology, 5*. doi:10.3389/fpsyg.2014.00980

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. Statistical science, 382-401.

Hsiao, Y.-Y., & Lai, M. H. C. (2018). The Impact of Partial Measurement Invariance on Testing Moderation for Single and Multi-Level Data. *Frontiers in Psychology, 9*. doi:10.3389/fpsyg.2018.00740

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics, 33*(2). doi:10.1214/009053604000001147

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (1st ed). Springer Texts in Statistics. doi:10.1007/978-1-4614-7138-7

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Jeon Minjeong, & Paul De Boeck. (2017). Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*. doi: 10.1037/met0000140.supp

Johnson, E. C., Meade, A. W., & DuVernet, A. M. (2009). The Role of Referent Indicators in Tests of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*. doi: 10.1080/10705510903206014

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika,*

36(4), 409–426. doi:10.1007/bf02291366

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

Kaplan, D., & Lee, C. (2015). Bayesian Model Averaging Over Directed Acyclic Graphs with Implications for the Predictive Performance of Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(3), 343–353. doi:10.1080/10705511.2015.1092088

Kaplan, D., & Lee, C. (2018). Optimizing Prediction Using Bayesian Model Averaging: Examples Using Large-Scale Educational Assessments. *Evaluation Review, 42*(4), 423–457. doi:10.1177/0193841x18761421

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*. doi: 10.1080/01621459.1995.10476572

Kelter, R. (2021). Bayesian model selection in theM-open setting — Approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator. *Journal of Mathematical Psychology*, 100, 102474. doi:10.1016/j.jmp.2020.102474

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing Measurement Invariance Using MIMIC. *Educational and Psychological Measurement*, *72* (3), 469–492. doi: 10.1177/0013164411427395

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

Lai, M. H. C., Kwok, O., Yoon, M., & Hsiao, Y.-Y. (2017). Understanding the Impact of

Partial Factorial Invariance on Selection Accuracy: An R Script. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(5), 783–799. doi:10.1080/10705511.2017.1318703

Lai, M. H., Richardson, G. B., & Mak, H. W. (2019). Quantifying the impact of partial measurement invariance in diagnostic research: An application to addiction research. *Addictive behaviors*, *94*, 50-56.

Lamers, S. M. A., Glas, C. A. W., Westerhof, G. J., & Bohlmeijer, E. T. (2012). Longitudinal evaluation of the mental health continuum-short form (MHC-SF): Measurement invariance across demographics, physical illness, and mental illness. *European Journal of Psychological Assessment*, *28* (4), 290–296. doi: 10.1027/1015-5759/a000109

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*(3), 662–668. doi:10.1037/0033-295x.112.3.662

Lei, H., Yao, S., Zhang, X., Cai, L., Wu, W., Yang, Y., … Zhu, X. (2016). Longitudinal Invariance of the Children's Depression Inventory for Urban Children in Hunan, China. *European Journal of Psychological Assessment, 32*(4), 255–264. doi:10.1027/1015-5759/a000195

Li, Y., Craig, B. A., & Bhadra, A. (2019). The Graphical Horseshoe Estimator for Inverse Covariance Matrices. *Journal of Computational and Graphical Statistics, 28*(3), 747–757. doi:10.1080/10618600.2019.1575744

Liang, X., & Luo, Y. (2019). A Comprehensive Comparison of Model Selection Methods for Testing Factorial Invariance. *Structural Equation Modeling: A Multidisciplinary Journal,* 1–16. doi: 10.1080/10705511.2019.1649983

Lin, L. C., Huang, P. H., & Weng, L. J. (2017). Selecting path models in SEM: A comparison

of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(6), 855-869.

Little, T. D. (1997). Mean and Covariance Structures (MACS) Analyses of Cross-Cultural Data: Practical and Theoretical Issues. *Multivariate Behavioral Research*, *32* (1), 53–76. doi: 10.1207/s15327906mbr3201{_}3

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn't be One. *Psychological methods*, *18*(3), 285.

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

Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*(428), 1535-1546.

Marcoulides, K. M., & Yuan, K. H. (2017). New Ways to Evaluate Goodness of Fit: A Note on Using Equivalence Testing to Assess Structural Equation Models. *Structural Equation Modeling*. doi: 10.1080/10705511.2016.1225260

Meade, A. W., & Bauer, D. J. (2007). Power and Precision in Confirmatory Factor Analytic Tests of Measurement Invariance. Structural Equation Modeling: A Multidisciplinary Journal, 14(4), 611–635. doi:10.1080/10705510701575461

Mellor, D., Vinet, E. V., Xu, X., Hidayah Bt Mamat, N., Richardson, B., & Román, F. (2015). Factorial invariance of the DASS-21 among adolescents in four countries. *European Journal of Psychological Assessment*, *31* (2), 138–142. doi: 10.1027/1015-5759/a000218

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

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian Structural Equation Models via

Parameter Expansion. *Journal of Statistical Software*, *85* (4), 1–30. doi: 10.18637/jss.v085.i04

Michalczyk, K., Maistädt, N., Worgt, M., Könen, T., & Hasselhorn, M. (2013). Age differences and measurement invariance of working memory in 5 to 12-year-old children. *European Journal of Psychological Assessment*, *29* (3), 220–229. doi:10.1027/1015-5759/a000149

Millsap, R. E. (2011). The Factor Model and Factorial Invariance. In R. E. Millsap (Ed), *Statistical Approaches to Measurement Invariance* (pp. 85 - 88). New York, NY: Routledge Taylor & Francis Group.

Millsap, R. E., & Kwok, O.-M. (2004). Evaluating the Impact of Partial Factorial Invariance on Selection in Two Populations. *Psychological Methods, 9*(1), 93–115. doi:10.1037/1082-989x.9.1.93

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In *Factor analysis at 100: Historical developments and future directions.* (pp. 131–152). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association, 83*(404), 1023–1032. doi:10.1080/01621459.1988.10478694

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

Muthén, B., & Asparouhov, T. (2013). *BSEM measurement invariance analysis.*

Muthén, B., & Asparouhov, T. (2017). Recent Methods for the Study of Measurement Invariance with Many Groups. *Sociological Methods & Research*, 004912411770148. doi: 10.1177/0049124117701488

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

Navarro, D.J. (2019). Between the Devil and the Deep Blue Sea: Tensions Between Scientific Judgement and Statistical Model Selection. *Comput Brain Behav 2*, 28–34 doi: 10.1007/s42113-018-0019-z

Owens, T. (1994). Two dimensions of self-esteem: Reciprocal effects of positive self-worth and self-deprecation on adolescent problems. *American Sociological Review, 59*(3), 391-407. doi:10.2307/2095940

Patalay, P., Deighton, J., Fonagy, P., & Wolpert, M. (2015). Equivalence of paper and computer formats of a child self-report mental health measure. *European Journal of Psychological Assessment*, *31* (1), 54–61. doi: 10.1027/1015-5759/a000206

Piironen, J., & Vehtari, A. (2016). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing, 27*(3), 711–735. doi:10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017, April). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. In *Artificial Intelligence and Statistics* (pp. 905-913). PMLR.

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics, 11*(2). doi:10.1214/17-ejs1337si

Pokropek, A., Schmidt, P., & Davidov, E. (2020). Choosing Priors in Bayesian Measurement Invariance Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(5), 750–764. doi:10.1080/10705511.2019.1703708

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-

project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology, 25* 111-164.

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92(437), 179–191. doi:10.1080/01621459.1997.10473615

Rensvold, R. B., & Cheung, G. W. (1998). Testing Measurement Models for Factorial Invariance: A Systematic Approach. *Educational and Psychological Measurement*, *58* (6), 1017–1034. doi: 10.1177/0013164498058006010

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving for the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management* (Vol. 1, pp. 25-50). Greenwich, CN: Information Age Publishing.

Rosenberg, M. (1979). *Conceiving the self*. New York, NY: Basic Books.

Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, ***48***(2), 1–36. https://www.jstatsoft.org/v48/i02/.

Shi, D., Song, H., DiStefano, C., Maydeu-Olivares, A., McDaniel, H. L., & Jiang, Z. (2019). Evaluating Factorial Invariance: An Interval Estimation Approach Using Bayesian Structural Equation Modeling. *Multivariate Behavioral Research, 54*(2), 224–245. doi: 10.1080/00273171.2018.1514484

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, 52*(4), 430–444. doi:10.1080/00273171.2017.1306432.

Stan Development Team (2020). "RStan: the R interface to Stan." R package version 2.19.3, http://mc-stan.org/.

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. doi: 10.1080/10705511.2021.1948335

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47*(2), 175–186. doi:10.1007/bf02296273

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

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

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

Vehtari A, Gabry J, Magnusson M, Yao Y, Bürkner P, Paananen T, Gelman A (2020). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 2.4.1, <URL:https://mc-stan.org/loo/>.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, *27*(5), 1413-1432.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Verhagen, A. J., & Fox, J. P. (2012). Bayesian tests of measurement invariance. *British*

Journal of Mathematical and Statistical Psychology, n/a–n/a. doi:10.1111/j.2044-8317.2012.02059.x

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383-401. doi: 10.1111/j.2044-8317.2012.02059.x

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. doi: 10.1016/j.jmp.2015.06.005

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. doi: 10.1016/j.jmp.2015.06.005

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review, 14*(5), 779-804. doi: 103758/bf03194105

Wagenmakers, E. J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology*. doi: 10.1016/j.cogpsych.2009.12.001

Watters, C. A., Taylor, G. J., Ayearst, L. E., & Michael Bagby, R. (2019). Measurement Invariance of English and French Language Versions of the 20-Item Toronto Alexithymia Scale. *European Journal of Psychological Assessment*, *35* (1), 29–36. doi: 10.1027/1015-5759/a000365

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the Meaning of Factorial Invariance and Updating the Practice of Multi-group Confirmatory Factor Analysis: A Demonstration with TIMSS Data. *Practical Assessment, Research & Evaluation*, *12*(3).

Yoon, M., & Kim, E. S. (2013). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods, 46*(4), 1199–

1206. doi:10.3758/s13428-013-0430-2

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*. doi: 10.1080/10705510701301677

Yuan, K. H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*. doi: 10.1037/met0000080

Zumbo, B. D., & Koh, K. H. (2005). Manifestation Of Differences In Item-Level Characteristics In Scale-Level Measurement Invariance Tests Of Multi-Group Confirmatory Factor Analyses. *Journal of Modern Applied Statistical Methods, 4*(1), 275–282. doi:10.22237/jmasm/1114907040