Improving the guidelines to conduct multigroup invariance test in Bayesian SEM

by

Esteban Montenegro-Montenegro, M.Ed.

A Dissertation

In

Educational Psychology

Submitted to the Graduate Faculty
of Texas Tech University in
Partial Fulfillment of
the Requirements for
the Degree of

DOCTOR OF PHILOSOPHY

Approved

Todd Little, PhD.
Chairperson

Jaeehoon Lee, PhD.
Co-chair

Mauricio Garnier-Villarreal, PhD.

Kwanghee Jung, PhD.

Mark Sheridan, Ph.D.
Dean of the Graduate School

May, 2020

ACKNOWLEDGEMENTS

The journey to finish this document has been long and sometimes bittersweet. To continue going I was always supported by many individuals who gave me encouragement, wisdom and affection. My parents, Miriam Montenegro and Luis Antonio Montenegro were from far away always attentive to provide love and positive words. Many hours on the computer talking to them helped me to find energy to move forward. I am thankful to my advisor Todd for always being patient and believe in me since the first day, thanks to Jaehoon Lee for your help in this wonderful journey, and finally but not less important I want to thanks Mauricio Garnier-Villarreal for sharing your knowledge and friendship.

The list of people is long but I have to be brief, nonetheless I don't want to exclude great supporters such as Daniel Bontempo who always gave me good ideas and advice, and David Johnson for your help in my academic career. Thanks to Sofia Brizuela for understanding my contradictions, hours of work and make the best company during this adventure. Finally, I know my brothers suffered because my distance but, with the same amount of positive feelings they were eager to see me happy.

CONTENTS

Texas Tech University, *Esteban Montenegro*, May 2020

ABSTRACT

The aim of the study was to evaluate goodness-of-fit measures in the context of Bayesian Structural Equation Modeling (BSEM) and invariance testing in multigroup models. Garnier-Villarreal and Jorgensen (2020) adapted several approximate fit measures usually applied in frequentist approach. They provided evidence of these adapted measures in single group models, however there was a lack of guidance on how to make decisions in invariance testing in Bayesian approach. I focused my simulation design to test the Bayesian adaptations of the indices: Comparative Fit Index (CFI), Tucker Lewis Index (TLI), Normed Fit Index (NFI), McDonals Centrality Index (Mc), and Gamma Hat ($\hat{\Gamma}$) index. The results showed that more conditions need to be added to find more evidence of the qualities of these measures. However, this study showed preliminary findings that support the implementation of the Bayesian CFI and Bayesian $\hat{\Gamma}$ in invariance testing. This work is part of a boarder effort to provide more evidence of appropriate fit measures in BSEM.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Measurement invariance is an analysis that aims to test the degree to which measurements show identical psychometric properties under different conditions Horn and McArdle (1992). These conditions can be longitudinal measurements (Little, 2013) or measurements conducted in different cultural groups (Little, 1997). According to Meade et al. (2008) the interest on measurement invariance has been increasing, given that there is more awareness of the importance of comparing equivalent measures.

The measurement invariance is nowadays more commonly tested implementing Confirmatory Factor Analysis (CFA) which is the measurement base of Structural Equation Modeling (SEM) and helps to evaluate the relationship between the observe variables (e.g. items in psychometrics) and the common latent factor that explains the observed variables (Brown, 2014; Kline, 2015).

The measurement invariance is tested by fitting several nested models that evaluate the psychometric equivalence of the items and the means and covariance structure of the latent factors (Cheung & Rensvold, 2002; Little, 1997, 2013; Meade et al., 2008). Thus, the constrained nested models represent a sequence of steps to get enough information to assume invariance in our measurement: first a configural model is estimated to assess the extent to which the proposed model fit the data between groups or overtime. In this model the objective is to evaluate if the same factorial structure can be utilized between different groups (e.g. sex, age group, cultural or ethnic group) or overtime (e.g. an intervention with repeated measures),

secondly a weak invariance (Little, 1997) model or metric invariance (Meredith, 1993) is estimated imposing constraints to factor loadings. In this step, we test the null hypothesis that all factor loadings are equivalent between groups or measurement occasions. This model is compared to the configural model where all loadings are freely estimated. As a third step, once weak invariance is tested and held, the intercepts of the manifest variables are constraint to be equal between groups. This is known as the strong invariance assumption or scalar invariance model. In this model, the null hypothesis of intercept equivalence between groups or overtime is tested (Cheung & Rensvold, 2002; Little, 1997, 2013; Meade et al., 2008). In this approach, every model has to hold the null hypothesis of equality of parameters to consider the model holds the tested assumption, once the proposed model holds strong invariance it is possible to test equality of latent variances, latent means and latent covariances (Little, 1997; Meredith, 1993)

In many cases, the nested models in invariance testing have been evaluated by performing a Likelihood Ratio-Test (LRT) (Schmitt & Kuljanin, 2008). However, this practice has been confronted by Cheung and Rensvold (2002) , and Meade et al. (2008). Cheung and Rensvold (2002) recommend to include the Comparative Fit Index (CFI)(Bentler, 1990), they also advised to consider a change of CFI (ΔCFI) larger than .01 as and important magnitude of misfit. Meade et al. (2008), revisited Cheung and Resnvold's study and conducted a second study that aimed to answer questions related to the power to detect non-invariant parameters, after conducting the study they recommend the cut-off ΔCFI no larger than 0.002 to accept the null hypothesis of invariance.

In Bayesian Structural Equation Modeling (BSEM) the guidelines for conducting

invariance testing are less clear compare to the frequentist approach described before. The most common practice in BSEM is to compare models using the deviance information criterion (DIC) which will be explained in detail in this manuscript. However, Garnier-Villarreal and Jorgensen (2020) adapted the TLI, CFI, RMSEA and gamma hat ($\hat{\gamma}$) to be used in BSEM as new peaces of information to evaluate the model fit. This innovation also allows to compare the ML and BSEM estimation by comparing these fit measures. The main aim of this study if to offer guidelines to evaluate the assumption of invariance, and test this new Bayesian fit measures under different conditions of misfit between groups. I expect to find that the cut-off values of ΔCFI suggested by Meade et al. (2008) could be also used in the Bayesian formulation.

1.1 Statement of the problem

Given the lack of literature addressing model fit indices in BSEM (Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018; Levy, 2011) and especially, only a few of substantial studies in BSEM invariance testing have been published. I aimed to create preliminary guidelines to conduct invariance testing in BSEM. Secondly, my goal is to test the new Bayesian adaptation of fit measures created by Garnier-Villarreal and Jorgensen (2020) in the context of invariance testing under different conditions of misfit, thirdly; I aim to compare the guidelines mentioned by Cheung and Rensvold (2002) and Meade et al. (2008) in MLE versus my findings in BSEM.

1.2 Bayesian Inference

Gelman et al. (2013) defines Bayesian inference as "the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations" (p.1). In their definition they are mentioning several concepts, and steps in Bayesian inference that need to be unpacked. Thus, it is necessary to start from the Baye's Rule, as stated by Gelman et al. (2013):

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{1.1}$$

Equation 1.1 states that the probability of parameter θ will be estimated given the data y , in Bayesian analysis probability statements are transcendental to understand the model that is going to be estimated. This is a model where the joint probablity distribution of θ and y are estimated, in simple words the Baye's rule estimates the product of two density functions or probability masses, they are the prior distribution $p(\theta)$ and the sampling distribution $p(y|\theta)$ also called the *likelihood of the data*.

As mentioned by Gelman et al. (2013), by using the conditional property showed in equation 1.1 we can estimate the posterior density as:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{1.2}$$

where $p(y) = \sum_\theta p(\theta)p(y|\theta)$, this is a sum of all possible values of θ, in the case of discrete distributions, but in continuous distributions we have to integrate: $\int p(\theta)p(y|\theta)d\theta$.

The Baye's Rule can also be expressed removing the factor $p(y)$ given that it does not depend on θ, and therefore y would be a constant that can be removed from the operation. This will yield the unnormlized posterior density (Gelman et al., 2013; Kaplan, 2014):

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{1.3}$$

In sum, Bayesian inference utilizes data to reallocate beliefs, it is in reality a mathematical formulation to reallocate and update beliefs based on data (Kruschke, 2010). Beliefs are represented by prior beliefs ($p(\theta)$) and are multiplied by the likelihood ($p(y|\theta)$) in order to get the posterior distribution ($p(\theta|y)$) which is a result that could be used to update or reallocate our previous belief.

The Bayesian statistical inference has an important distinction regarding the nature of the unknown parameters, in frequentist inference the unknown parameters (θ) are seen as fixed values, whereas in Bayesian inference the unknown parameters are considered as a random variable with a probability distribution that represents the uncertainty about the true value of θ (Kaplan, 2014).

1.2.1 The prior distribution

Bayesian inference requires the inclusion of beliefs to estimate the posterior distribution of the unknown parameters. Kaplan (2014) remarks that priors represent the knowledge that previous studies collect or show in a specific matter. The idea is to incorporate previous information that we are always revisiting when planing our designs and interventions, to some extent we always have previous information and hypothesis in mind, these hypothesis and previous information can

be included in the model for Bayesian inference. However, previous information can be also sparse and unclear, or the researcher does not have enough knowledge regarding the specific matter, in this case the Bayesian approach can also help to account for the uncertainty , it is also possible to test models with more or less uncertainty and make decision based on the posterior distribution.

There are two main types of priors, a prior can be *noninformative* when we don't have enough evidence or strong evidence, in this case we can specify a *vague* prior distribution, for instance is usual to use a uniform distribution as a prior to quantify the level of ignorance (Kaplan, 2014). This implementation of the uniform distributions is known as the *principle of insufficient reason* coined by Laplace who claimed that a uniform distribution is appropiate when nothing is known about θ (Gelman et al., 2013). The *informative prior* is another type of prior, it is used when the level of uncertainty is lower and enough information is available to believe that the parameter has a specific distribution where all values are closer, for example, to the expected value of the distribution, for instance we could believe that the distribution of the slope in a regression follows a normal distribution with mean zero and very small variance let's say .001 ($\theta \sim N(0, .001)$). This prior is a very informative prior because it represents the believe that the posterior values will be all closer to the mean. Informative priors can also have a strong influence on the posterior distribution if they are very informative, especially when the number of observations is small (Kruschke, 2010)

In this study strong or informative priors will be relevant for invariance testing in BSEM, by implementing the approach so-called *approximate Measurement Invariance* Muthén and Asparouhov (2013), this implementation of strong priors

will be explained in the following sections.

1.2.2 MCMC estimation

In Bayesian inference it is relevant to summarize the posterior distribution, while the frequentist approach relies mainly on Maximum Likelihood estimation (ML), this means the derivation of point estimates that have optimal asymptotic properties (Kaplan, 2014). In Bayesian analysis the inference starts from estimating elements of the posterior distribution (e.g. the expected a posteriori or maximum a posteriori, posterior intervals), however this calculations require to estimate expectations that are the result of integrating, but this estimations can be difficult to perform in complex and high-dimensional models that include multiple parameters and integrals. The problem of complexity is solve in Bayesian inference by drawing samples from the posterior distribution and summarize the obtained samples(Kaplan, 2014). This approach is called *Mote Carlo integration* and it is based on the "first drawing T samples of the parameters of interest $\{\theta_t, t = 1, \ldots, T\}$ from the posterior distribution $p(\theta|y)$ and approximating the expectation by" (Kaplan, 2014, p. 66) :

$$E[p(\theta|y)] \approx \frac{1}{T}\sum_{t=1}^{T} p(\theta|y). \tag{1.4}$$

Equation 1.4 is more accurate as T increases following the law of large numbers, assuming that the samples are independent. However, the assumption of independence can be relaxed using a Markov chain. According to Robert and Casella (2010) and Gelman et al. (2013), a Markov chain $\{X^{(t)}\}$ is a sequence of dependent random variables $X^0, X^1, X^2, \ldots, X^t, \ldots$ such that the probability

distribution of X^t given the past variables depends only on $X^{(t-1)}$. This conditional probability is known as the transition kernel or Markov kernel K, which is (Robert & Casella, 2010, p. 168):

$$X^{(t+1)|}|X^0, X^1, X^2, \ldots, X^t \sim K(X^t, X^{(t+1)}). \tag{1.5}$$

Thus, a simple random walk Markov chain satisfies the condition:

$$X^{t+1} = X^t + \epsilon_t.$$

where $\epsilon_t \sim N(0, 1)$ and therefore, the Markov kernel $K(X^t, X^{(t+1)})$ corresponds to a $N(X^{(t)}, 1)$ density. The Markov Chain Monte Carlo (MCMC) simulations has a strong stability property, thus a stationary probability distribution by constructing those chains, there is a probability distribution such that if $X^{(t)} \sim f$, then $X^{(t+1)} \sim f$. Given this property, we can expect the chain to converge to its stationary distribution $p(\theta|y))$ (posterior distribution) after several iterations where the approximate distributions are improved at each step in the simulation (Gelman et al., 2013; Kaplan, 2014; Robert & Casella, 2010) .

In sum, Markov chain simulation and specially MCMC is used when is computationally difficult to sample θ from $p(\theta|y)$, in this case the idea is to sample iteratively and at each step of the process is expected to draw values from a distribution that becomes closer to $p(\theta|y)$. Therefore, the aim is to create a Markov chain simulation whose stationary distribution is the posterior distribution $(p(\theta|y))$ and to run the simulation long enough that the distribution of the draws is approximately similar to the the stationary distribution. (Gelman et al., 2013,

p. 276).

There are different samplers or Markov chain algorithms that can be implemented, however the most commonly used are the *Gibbs sampler* and the *Metropolis-Hastings Algorithm*. The latter is explained in the following section.

1.2.2.1 Metropolis-Hastings Algorithm

The Metropolis-Hastings Algorithm is also called a random walk algorithm that draws values through the parameter space starting at some arbitrary point with an acceptance/rejection rule to converge . This algorithm is frequently used in complex models and can be more efficient than Gibbs sampler to solve complex dimensional problems (Gelman et al., 2013; Kruschke, 2010).

The algorithm in this case will select a value x from the proposal distribution (*jumping distribution*) that we can call $q(.|y_t)$, then the algorithm evaluates the selected value with a probability (Kruschke, 2010, p. 68):

$$p(x, y) = min\left\{1, \frac{p(x)q(y|x)}{p(y)q(x|y)}\right\} \tag{1.6}$$

In Equation 1.6 the numerator is the probability of the candidate value and the denominator is the probability of the current value. If the result of this ratios is 1.0 the algorithm accepts the candidate value, otherwise the algorithm will move to the next value or stay in the current value. To make a decision, the algorithm draws a value from a uniform distribution ($U(0, 1)$) , if the sample number is between 0 and $p(y, x)$ then the algorithm moves to the next value.

In simple words, we can imagine to have a distribution $P(\theta)$, in a multidimensional continuous space, the main point is to generate sample values

from this multidimensional space, thus we can estimate the value of $P(\theta)$ for any candidate value of θ. Sample values from the target distribution are estimated by the random walk in the multidimensional space. The random walk starts at a random point specified by the user, this initial location should be someplace where $P(\theta)$ is not zero. The random walker proposes to move a new place in each step and then deciding whether or not to accept the proposed move (Kruschke, 2010, p. 157).

1.2.2.2 Chains convergence

It is frequent to generate several MCMC chains to draw values from the posterior distribution. However, the chains might draw values from different areas in the posterior distribution but eventually all the chains will converge in the long run. But this is not always the case and it is important to evaluate and measure if the chains have converged in the same distribution. Trace plots and autocorrelation function plots (ACF) can be used to diagnose the convergence of the Markov chains, these methods can be also complemented with the potential scale reduction factor (PSRF) also known as $\hat{R}$ (Brooks & Gelman, 1998). This statistic takes into account the within-chain variance and between-chain variance, in simple words the idea is to evaluate if the variance of the target distribution was overestimated or underestimated, values close to 1.0 show convergence of the chains. (Brooks & Gelman, 1998).

1.3 Bayesian Structural Models and Invariance Testing

Most of the literature focuses on estimating covariance structure SEM models, this approach fits the covariance structure under the proposed model to the sample covariance matrix. This approach works properly under normal distributions and

large sample sizes, however complex models and small sample sizes make the estimation difficult to achieved (Lee & Song, 2012).

The co-variance structure in SEM can be modeled following the well know extension of the exploratory factor analysis: the Confirmatory Factor Analysis, which is the measurement model in SEM (Kline, 2015; Lee, 2007):

$$x = \Lambda\xi + \epsilon, \tag{1.7}$$

where $\Lambda(p \times q)$ is a matrix of factor loadings, $\xi(q \times 1)$ is a random vector of latent common factors and $\epsilon(p \times 1)$ is a random vector of error measurements, it is also called latent unique factors or residuals. In this model, ξ is distributed as $N[0, I]$, and ϵ is distributed as $N[0, \Psi_\epsilon]$, where Ψ_ϵ is a diagonal matrix, and ξ is uncorrelated with ϵ Lee (2007, p.15). CFA models allow the correlation between latent common factors thus ξ is distributed as $N[0, \phi]$ with a positive definite covariance matrix ϕ, and ξ is assumed to be independent with ϵ, this will lead to the structural model:

$$\Sigma = \Lambda\Phi\Lambda^T + \Psi_\epsilon \tag{1.8}$$

In equation 1.8, different constraints are possible in order to identify the model, also elements in Λ, Φ and Ψ_ϵ can be fixed, the fixed parameters are in this model values that represent hypothesis that will be tested in the model, for instance a value fixed as $\lambda_{kh} = 0$ means that the hth factor does not explain the kth observed variable. The model also comprises the specification of number of latent common factors and their correlation (Φ) (Lee, 2007).

The CFA model comprises measurement equations such as:

$$x_1 = \Lambda_1 \eta + \epsilon_1 \tag{1.9}$$

$$x_2 = \Lambda_2 \xi + \epsilon_2, \tag{1.10}$$

$$\tag{1.11}$$

Where $x_1(r \times 1)$ and $x_2(s \times 1)$ are manifest variables that will be added as indicators for η and ξ. The matrices $\Lambda_1(r \times q_1)$ and $\Lambda_2(s \times q_1)$ are loading matrices, the random vectors $\epsilon_1(r \times 1)$ and $\epsilon_2(s \times 1)$ represent error measurements. These errors are assumed to be uncorrelated with η and ξ, the distributions of the error is normal with zero means. Given the random vectors x_1 and x_2 , the measurement equations account for the variances that the manifest variables have in common allowing the estimation of common factors, this also opens the possibility of estimating a structural model between la latent common factors by solving the next equation:

$$\eta = \Pi \eta + \Gamma \xi + \delta \tag{1.12}$$

Where $\eta(q_1 \times 1)$ is an endogenous random vector of latent variables and $\xi(q_2 \times 1)$ is an exogenous random vector of latent variables, $\Pi(q_1 \times q_1)$ and $\Gamma(q_1 \times q_2)$ are matrices of regression coefficients between η and ξ. The vector $\delta(q_1 \times 1)$ comprises error measurements or residuals. The assumption is that ξ is uncorrelated with δ, and the means are zero for these two vectors.

Bayesian SEM follows the same model structure described up to this point, however the model specification has to take into account adequate priors to improve

model convergence. Following Lee (2007), let M be a SEM with a vector of unknown parameters θ, and let Y be the observed data with a sample size n. In BSEM θ is considered random whereas from a frequentist approch it will be fixed, thus θ has a prior distribution, for instance $p(\theta|M)$ or $p(\theta)$ for the sake of simplicity.

The joint probability of Y and θ under M is $p(Y, \theta|M)$ which can be also called the likelihood of the model as explained in Chapter 1.2, these distribution added together will result in the estimation of the posterior distribution $p(\theta|Y, M)$ as:

$$log(p\theta|Y, M) \propto logp(Y|\theta, M) + logp(\theta) \tag{1.13}$$

In equation 1.13 the term $p(Y|\theta, M)$ can be seen as the likelihood function. As explained before, sample size can affect the likelihood, large sample sizes can make the likelihood larger therefore, it dominates $logp(\theta)$)(prior).

Lee (2007) states that conjugate priors are frequently added in Bayesian SEM, and they recommend to utilize conjugate priors. For instance, they include inverse gamma distributions on variance parameters, inverse Wishart distributions on covariances of exogenous latent variables, and normal distributions on other parameters. However, Lee (2007) keep the assumption that the manifest variables covariance matrix Θ and the latent covariance matrix Ψ are diagonal matrix, this assumption does not allow to fit more realistic models. Merkle and Rosseel (2018) incorporated a method in `blavaan` package to overcome this issue.

1.3.1 Bayesian estimation

Lee (2007) illustrates the estimation of the BSEM from a CFA model, for $i = 1, \ldots, n$:

$$y_i = \Lambda\omega_i + \epsilon_i \tag{1.14}$$

In equation 1.14 y_i is a $p \times 1$ observed random vector, Λ is factor loading matrix, ω_i is a vector of factor scores and ϵ_i is a random vector of error measurements independent of ω_i. The distribution of ϵ_i is assumed to be normal ($\epsilon \sim N(0, \Psi)$) where Ψ could be a diagonal matrix[1], ω_i is also assumed to be normally distributed ($\omega \sim N(0, \Phi)$) where Φ is a positive definite covariance matrix.

The Gibbs sampler will follow this process:

1. Generate $\Omega^{(j+1)}$ from $p(\Omega|\Psi_\epsilon^{(j)}, \Lambda^{(j)}, \Phi^{(j)}, Y)$
2. Generate $\Psi_\epsilon^{(j+1)}$ from $p(\Psi_\epsilon|\Omega^{(j+1)}, \Lambda^{(j)}, \Phi^{(j)}, Y)$
3. Generate $\Lambda^{(j+1)}$ from $p(\Lambda|\Omega^{(j+1)}, \Psi^{(j+1)}, \Phi^{(j+1)}, Y)$
4. Generate $\Phi^{(j+1)}$ from $p(\Phi|\Omega^{(j+1)}, \Psi^{(j+1)}, \Lambda^{(j+1)}, Y)$,

Where, $Y = (y_1, \ldots, y_n)$ is the observed data matrix, Ω is a matrix of latent factor scores which will be treated as as hypothetical missing data, estimated from the augmented observed data Y with Ω in the posterior analysis.

1.3.2 Data Augmentation

As mentioned before, the estimation of factor scores in BSEM is conducted by the implementation of data augmentation (Merkle, 2011; Merkle & Rosseel, 2018), which is a strategy utilized in multiple imputation (MI). Data Augmentation (DA) algorithm comprises imputing iteratively missing data and sampling data from a

[1]This matrix could have covariances following the work of Merkle and Rosseel (2018)

Bayesian model via Markov Chain Montecarlo, the results are summarized from the posterior distribution (Merkle, 2011). In order to treat the missing data patterns with MI, the assumption of missing at random should be accepted (Schafer, 1997). This assumption states that the missing patterns depend on the observed data and not on the missing data, this could be expressed as:

$$P(M|Y_{obs}, Y_{mis}, \xi) = P(M|Y_{obs}, \xi) \tag{1.15}$$

Where Y is a data matrix of n observations measured on p variables. The missing part of Y is label as Y_{mis} and the observed part as Y_{obs}. The vector ξ contains parameters that depend on the missing data mechanism. The M represents a $n \times p$ matrix with elements equals to zero if the element in Y is observed otherwise it is equal to 1 for missing elements. Thus, in equation 1.15 we can see that the probability of a missing observation depends on Y_{obs} but not on Y_{mis}.

The probablity of missing can be also independent of the observations and the missing pattern itself by holding the assumption of Missing Completely at Random (MCAR), the formal expression can be written as (Merkle, 2011):

$$P(M|Y_{obs}, Y_{mis}, \xi) = P(M|\xi) \tag{1.16}$$

In equation 1.16 the probability of a missing element (represented as 1 in vector M) is not dependent on the observed values nor the missing itself. It is a completely random process.

These missing data mechanisms are behind the Bayesian estimation with data augmentation which is a well known approach to handle latent variables (Song &

Lee, 2012).

Following the same treatment of latent factor scores as missing values we can consider the following steps based on Gibbs sampling (Merkle, 2011; Song & Lee, 2012):

1. Draws $Y_{mis}^{(t+1)}$ from $p(Y_{mis}|Y_{obs}, \theta^{(t)})$;
2. Draws $\theta^{(t+1)}$ from $p(\theta|Y_{obs}, Y_{mis}^{(t+1)})$.

The step 1 can be called "the imputation step" and step 2 can be considered the "posterior step"(Merkle, 2011, p. 261). The sampled θ values converge the observed-data posterior distribution $p(\theta|Y_{obs})$. The main reason to implement these steps

1.3.3 Invariance testing in BSEM

Muthén and Asparouhov (2013) proposed the approximate measurement invariance analysis, where informative priors are included instead of equating parameters between groups. Thus, the exact zero-constraints can be replaced with approximate zero constraints. Differences, for instance in items' intercepts can be estimated with the "wiggle room" proposed by Muthén and Asparouhov (2012) , in this approach very small differences are allowed and thus finding a compromise between zero and no constraints (Chiorri et al., 2014; Van De Schoot et al., 2013).

The prior distribution has the most important role in approximate MI, specifically a strongly informative prior is added in order to estimated the posterior of the difference between parameters. Because priors with small variances will pull the posterior distribution of the difference between parameters to be approximately

zero, this approach is more realistic than the assumption of exact zero difference in MLE. (Muthén & Asparouhov, 2013; Van De Schoot et al., 2013).

This approach has shown to be a important alternative compare to partial invariance. As pointed out by Van De Schoot et al. (2013) , several studies have claimed that partial invariance is a sufficient condition to test differences between latent means and get unbiased estimated latent means (Byrne et al., 1989; Muthén & Christoffersson, 1981), however approximate MI performs better at getting unbiased estimates compare to partial invariance, showing that approximate MI is actually a suitable approach when there are multiple differences larger than zero between several parameters, it also helps to get models that fit the data properly compare to partial invariance (Van De Schoot et al., 2013).

Despite of measurement invariance popularity there is not published articles offering guidelines or cutoff points to test invariance in Bayesian SEM (BSEM). It is possible to find applied articles testing BSEM invariance implementing a Bayes factor comparison, or the posterior predictive p-value (PPP) (Barendse et al., 2014; De Bondt & Van Petegem, 2015; Muthén & Asparouhov, 2013). In addition, several articles report comparisons using the DIC (Bujacz et al., 2014). However, none of these articles provide a precise simulation to get the most efficient and unbiased statistic to detect the lack of invariance, also these investigations do not provide how large should be the difference in these fit measures to hold the assumption of invariance.

1.3.4 Model fit indices

There are well known fit indices in frequentist SEM approach, Comparative Fit Index (CFI), Tucker-Lewis (TLI), Root Mean Square Error, and the χ^2 test statistic

are among the most cited indices in SEM literature. However, fit indices for BSEM have not been extensively explored (Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018; Levy, 2011). Currently, two studies explored alternative Bayesian formulations, Hoofs et al. (2018) proposed a Bayesian estimation for the RMSEA named BRMSEA, and Garnier-Villarreal and Jorgensen (2020) expand the formulation of Hoofs et al. to obtain a Bayesian CFI, TLI and $\hat{\Gamma}$. In this section I will review the fit indices aforementioned from the frequentist perspective, then I will explain the Bayesian adaptations elaborated by Garnier-Villarreal and Jorgensen, and Hoofs et al.

1.3.4.1 The χ^2 Test

The χ^2 test is based on the discrepancy function, this discrepancy evaluates how distant is the model from the observed covariance matrix. In a formal definition, the discrepancy function compare the sample covariance matrix S to the model-implied covariance matrix $\Sigma(\hat{\theta})$. The maximum likelihood discrepancy function is expressed as (Garnier-Villarreal & Jorgensen, 2020):

$$F_{ML} = log|\hat{\Sigma}| - log|S| + trace(S\hat{\Sigma}) - p \tag{1.17}$$

In 1.17 p is the number of variables in the model. This can be also modified to include the mean a covariance structure (MACS):

$$F_{ML} = log|\hat{\Sigma}| - log|S| + trace(S\hat{\Sigma}) - p + (\bar{x} - \hat{\mu})^T\hat{\Sigma}^{-1}(\bar{x} - \hat{\mu}). \tag{1.18}$$

Where $\bar{x}$ is the vector of sample means and $\hat{\mu}$ is the vector of model-implied means. The result follows a central chi-square distribution with degrees of freedom

equal to that of the model, thus the statistic is estimated as $\chi^2_{ML} = N(F_{ML})$ or $\chi^2_{ML} = N - 1(F_{ML})$ where N represents the number of observations (Kline, 2015).

We can also estimate the χ^2_{ML} of a model by estimating the multivariate log-likelihood function(l) and compare the hypothesized model versus a saturated model, first we estimate:

$$l_n = -\frac{p}{2}log(2\pi) - \frac{1}{2}log|\Sigma| - \frac{1}{2}(y_n - \mu)^T\Sigma^{-1}(y_n - \mu). \tag{1.19}$$

Where n is the n^{th} observation of the vector y_n. Then, after plugging each n^{th} value and adding the values we obtain the log-likelihood of the hypothesized model,

$$l_h = \sum_{n=1}^{N} l_n \tag{1.20}$$

The same steps can be followed to estimate the saturated model, the only change to me made is to plug the observed sample statistic $\bar{x}$ and S instead of μ and Σ. Finally, we can compare two models, the hypothesized model and the saturated model utilizing log-likelihood ratio test (LRT):

$$\chi^2_{ML} = -2(l_H - l_S) \tag{1.21}$$

The result of the LRT is χ^2 distributed with $df = p - q$ where p is the number of non redundant sample moments and q is the number of parameters to be estimated.

In sum, the χ^2 statistic evaluates the hypothesis of exact model fit, this means that H_0 states that the model proposed represents the true data-generating process (Garnier-Villarreal & Jorgensen, 2020; Kline, 2015). However, the χ^2 statistic is very sensitive to detect misfit especially when sample size is large (Bentler, 1990;

Kline, 2015), therefore it is recommended to complement the evaluation of the model including several fit indices.

1.3.4.2 Root Mean Square Error of Approximation (RMSEA)

The RMSEA is noncentrality fit measure, it is also considered as an absolute fit index, where a value of zero indicates the best results. This statistic favors models with more degrees of freedom or models with large sample sizes with lower values of RMSEA (Kline, 2015).

The formal expression is,

$$\hat{\epsilon} = \sqrt{max\left[0, \frac{\hat{\lambda}}{df(N)}\right]} = \sqrt{max\left[0, \frac{\chi^2_{ML} - df}{df(N)}\right]} = \sqrt{max\left[0, \frac{\chi^2_{ML}}{df} - \frac{1}{N}\right]} \quad (1.22)$$

In equation 1.22 the model misfit is expressed as an average across the number of restrictions. To represent this average per degrees of freedom (df) in the metric of the discrepancy function, $\hat{\lambda}$ is divided by N, and the degrees of freedom (df).

The RMSEA has a known sampling distribution that makes possible to estimate confidence intervals to test null hypothesis about specific population values of RMSEA (Browne & Cudeck, 1992; MacCallum et al., 2006).

Values $\hat{\epsilon} < .05$ are considered evidence of close fit, $.05 < \hat{\epsilon} < .08$ indicates reasonable fit, $.08 < \hat{\epsilon} < .10$ indicates mediocre fit, and $\hat{\epsilon} > .10$ is an unacceptable fit for a model (Browne & Cudeck, 1992).

1.3.4.3 Gamma hat ($\hat{\Gamma}$)

The McDonald's centrality index (Mc) is a measure that also takes $\hat{\lambda}$ to divide it by N , but it is additionally exponentiated to express misfit in terms of likelihood, values close to 1.0 indicate excellent model fit (McDonald, 1989):

$$Mc = e^{-\frac{1}{2}(\frac{\hat{\lambda}}{n})} \tag{1.23}$$

A modification to Mc was proposed introducing the number of variables p:

$$\hat{\Gamma} = \frac{p}{p + 2\frac{\hat{\lambda}}{n}} \tag{1.24}$$

Similar to Mc, values of gamma hat ($\hat{\Gamma}$) close to 1.0 indicate a better fit.

1.3.4.4 Incremental fit indices

The Tuker-Lewis Index (TLI) and the Comparative Fit Index (CFI) are two incremental fit indices, both indices range from 0 to 1.0, values close to 1.0 are interpreted as a better fit. The CFI and TLI evaluates the amount of departure from close fit for the hypothesized model versus the independent model or null model (Kline, 2015).

Tucker and Lewis (1973) proposed the TLI - also known as nonnormed fit index (NNFI)- as a reliability index for exploratory factor analysis, the aim was to create an index that helped to select the number of factor.

$$TLI = NNFI = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_H^2}{df_H}}{\frac{\chi_0^2}{df_0} - 1} \tag{1.25}$$

Equation 1.25 is the estimation of the TLI where H stands for hypothesized

model and the "0" subscript represents the null or the independent model. The TLI is not bound to values between 0 and 1.0 because it can assume values higher than 1.0 when $x_H^2 < df_H$(Garnier-Villarreal & Jorgensen, 2020).

The CFI is similar to TLI, it was proposed by Bentler (1990) and it is normed to fall between 0 and 1 (Garnier-Villarreal & Jorgensen, 2020; Kline, 2015):

$$CFI = \frac{max(0, \hat{\lambda}_0) - max(0, \hat{\lambda}_H)}{max(0, \hat{\lambda}_0)} = 1 - \frac{max(0, \hat{\lambda}_H)}{max(0, \hat{\lambda}_0)} \tag{1.26}$$

1.3.4.5 Bayesian fit indices

Most of the software, and published articles report only measures of overall fit such as the posterior predictive p value (PPP, Gelman et al. (1996)). Also, for model comparasion the most common index is the deviance information criterion (DIC, Spiegelhalter et al. (2002a)). However, Garnier-Villarreal and Jorgensen (2020) revisited the BRMSEA proposed by Hoofs et al. (2018) in order to create a Bayesian CFI, TLI and $\hat{\Gamma}$ estimation. The formulation of these Bayesian fit indices include aspects of the PPP and the effective number of parameters used to calculate DIC.

1.3.4.5.1 Posterior Predictive Checks

The posterior predictive checks were proposed by Gelman et al. (1996) and it aims to detect if the model fits the data properly, if the model fits, then a set of replicated data generated under the model should look similar to observed data (Gelman et al., 2013). The basic idea is to draw simulated values from the joint posterior predictive distribution of replicated data and after that, compare the new samples to the observed data, any difference can be evidence of misfit in the model (Gelman et al., 2013).

To detect any misfit in a model, a discrepancy function can be estimated to evaluate to what extent a feature of the observed data differs from its expected value given the model parameters at iteration i of a Markov chain. If the model does not fit the data, the value of discrepancy function will be large for the observes data D_i^{obs} (Garnier-Villarreal & Jorgensen, 2020).

Random samples are drawn from the posterior predictive distribution, at the same iteration i of the Markov chain, and because the replicated data are consistent with the model, the replicated data (D_i^{rep}) represents sampling error. Thus, if the model reflects the true-data-generating process, then $P(D^{obs} > D^{rep})$. This is known as the Posterior Predictive p-value (PPP), the PPP estimates the proportion of $i = 1, 2, ...I$ samples from the posterior for which $D_i^{obs} > D_i^{rep}$.

The posterior predictive checking can be used to estimate a standard likelihood-ratio chi-square statistic of an H_0 model against an unrestricted H_1 . The p value in this case can indicate poor or adequate fit, a PPP value around .5 is considered an excellent- fitting model (Muthén & Asparouhov, 2012) .

1.3.4.5.2 Deviance Information Criterion (DIC)

Spiegelhalter et al. (2002b) proposed the DIC as a Bayesian variation of Akaike's information criterion (AIC), similarly the DIC takes into account a penalty by the number of parameters (p) , however in Bayesian we need to estimate the effective number of parameters (pD) from the posterior distribution by calculating the deviance using each vector of parameters θ_i sampled from the posterior distribution during MCMC estimation. As a result, we will get a posterior distribution of the deviance with posterior mean $\bar{D}$. The deviance of different point-estimates ($D(\bar{\theta})$) can be also calculated using $\hat{\mu}$ and $\hat{\Sigma}$ implied by the posterior mean of the

parameters $\hat{\theta}$, the difference between $\bar{D}$ and $D(\bar{\theta})$ is called the effective number of parameters (Garnier-Villarreal & Jorgensen, 2020; Spiegelhalter et al., 2002b) :

$$pD_{DIC} = \bar{D} - D(\bar{\theta}) \tag{1.27}$$

Expression 1.27 can also be estimated in a log-likelihood metric:

$$pD_{DIC} = -2 \times [\bar{l}_H - l_H(\bar{\theta})] = -2 \times \left[\frac{1}{I} \sum_{i=1}^{I} log[P(Y|\hat{\theta}_i)] - log[P(Y|\bar{\theta})] \right] \tag{1.28}$$

1.3.4.5.3 Widely Applicable Information Criterion (WAIC)

The WAIC (Watanabe, 2010) has a similar formulation compared to the DIC describe before. As stated by Merkle and Rosseel (2018).

$$WAIC = -2lppd + 2efp_{WAIC}, \tag{1.29}$$

The first term represents the log-likelihoods of observed data and the second term is the effective number of parameters. The first term, called the log pointwise predictive density of the observed data (lppd) is estimated as:

$$lppd = \sum_{i=1}^{n} log \left(\frac{1}{S} \sum_{S=1}^{S} f(y_i|\theta^S) \right) \tag{1.30}$$

Where S is the number fo posterior draws and $f(y_i|\theta^S)$ is the density of observation i with respect to the parameter sampled at iteration s.

The effective number of parameter (efp_{WAIC}) is calculated as:

$$efp_{WAIC} = \sum_{i=1}^{n} var_s(logf(y_i|\theta)) \tag{1.31}$$

In equation 1.31 a separate variance is estimated for each observation i across the S posterior draws.

1.3.4.5.4 Leave-One-Out cross-validation (LOO)

The LOO measures the predictive density of each observation holding out one observation at the time and use the rest of the observations to update the prior. This estimation is calculated via (Merkle & Rosseel, 2018; Vehtari et al., 2015):

$$LOO = -2\sum_{i=1}^{n} log\left(\frac{\sum_{s=1}^{S} w_i^s f(y_i|\theta^s)}{\sum_{s=1}^{s} w_i^s}\right) \tag{1.32}$$

Where w_i^s are sampling weights taken from the relative magnitude of individual i density function across the S posterior samples.

LOO also requires the estimation of efective number of parameters that can be calculated taken the *lppd* term from WAIC:

$$efp_{LOO} = lppd + LOO/2 \tag{1.33}$$

1.3.4.5.5 Bayesian RMSEA (BRMSEA)

The BRMSEA was proposed by Hoofs et al. (2018), it incorporates the differences in discrepancies of observed and replicated data ($D_i^{obs} - D_i^{rep}$) at each iteration in the Markov chain similar to posterior model checks. Similarly, the BRMSEA includes the effective number of parameters (pD_{DIC}) as showed in equation 1.28, it also accounts for model complexity by adding $p * -pD$ to the

estimation(Garnier-Villarreal & Jorgensen, 2020; Hoofs et al., 2018):

$$BRMSEA_i = \sqrt{max\left[0, \frac{(D_i^{obs} - D_i^{rep}) - (p * -pD)}{(p * -pD) \times N}\right]} \tag{1.34}$$

The purpose of the BRMSEA is to complement the model evaluation along with the PPP given that the PPP tends to reject all models with even minor misspecifications when sample size is large. The BRMSEA also follows similar cutoff points compare to the frequentist RMSEA (Hoofs et al., 2018).

Garnier-Villarreal and Jorgensen (2020) proposed a modification in equation 1.34 to obtain a fit index that behaves like its frequentist counterpart by esimating the deviance evaluated at the posterior mean:

$$BRMSEA^{DevM} = \sqrt{max\left[0, \frac{(D_i^{obs} - pD) - (p^* - pD)}{(p^* - pD) \times N}\right]} \tag{1.35}$$

$$= \sqrt{max\left[0, \frac{D_i^{obs} - p^*}{(p^* - pD) \times N}\right]} \tag{1.36}$$

In Equation 1.35 D_i^{rep} is replaced by pD, and the two occurrences of pD cancel out, resulting in a numerator that expresses misfit as the discrepancy at iteration i rescaled by the number of observed sample moments (Garnier-Villarreal & Jorgensen, 2020).

1.3.4.5.6 Bayesian CFI, TLI and $\hat{\Gamma}$

Garnier-Villarreal and Jorgensen (2020) proposed in addition to BRMSEA a Bayesian estimation for CFI, TLI and $\hat{\Gamma}$. These new fit measures replace the maximum likelihood chi-square(χ^2_{ML}) with $(D_i^{obs} - pD)$ and df with ($p^* - pD$). In

addition, the parameter $\hat{\lambda}$ is replaced by $(D_i^{obs} - pD) - (p^* - pD) = D_i^{obs} - p^*$.

The formal estimation for these fit indices can be expressed as:

$$\text{BMc}_i^{\text{DevM}} = e^{-\frac{1}{2N}[(D_i^{obs} - pD) - (p^* - pD)]} = e^{-\frac{1}{2N}(D_i^{obs} - p^*)}. \tag{1.37}$$

$$\text{B}\hat{\Gamma}_i^{\text{DevM}} = \frac{p}{p + \frac{2}{N}[(D_i^{obs} - pD) - (p^* - pD)]} = \frac{p}{p + \frac{2}{N}(D_i^{obs} - p^*)}. \tag{1.38}$$

$$\text{B-}\hat{\Gamma}_{adj,i}^{\text{DevM}} = 1 - \frac{p^*}{p^* - pD}(1 - \text{B-}\hat{\gamma}_i^{\text{DevM}}). \tag{1.39}$$

$$\text{BTLI}_i^{\text{DevM}} = \text{BNNFI}_i^{\text{DevM}} = \frac{\frac{D_{0,i}^{obs} - pD_0}{p^* - pD_0} - \frac{D_{\text{H},i}^{obs} - pD_{\text{H}}}{p^* - pD_{\text{H}}}}{\frac{D_{0,i}^{obs} - pD_0}{p^* - pD_0} - 1}. \tag{1.40}$$

$$\text{BNFI}_i^{\text{DevM}} = \frac{(D_{0,i}^{obs} - pD_0) - (D_{\text{H},i}^{obs} - pD_{\text{H}})}{D_{0,i}^{obs} - pD_0}. \tag{1.41}$$

$$\text{BCFI}_i^{\text{DevM}} = 1 - \frac{(D_{\text{H},i}^{obs} - pD_{\text{H}}) - (p^* - pD_{\text{H}})}{(D_{0,i}^{obs} - pD_0) - (p^* - pD_0)} = 1 - \frac{D_{\text{H},i}^{obs} - p^*}{D_{0,i}^{obs} - p^*}. \tag{1.42}$$

Following the superscripts and subscripts from (Garnier-Villarreal & Jorgensen, 2020) , the superscript "DevM" indicates that the observed deviance at iteration i in the Markov chain is rescaled using the effective number of parameters (pD). The hypothesized model have an "H" subscript, the null model is represented with a "0" subscript.

CHAPTER 2
LITERATURE REVIEW

This section describes the main studies that aim to create new fit measures for BSEM, however up to date there is not sufficient literature about model evaluation in BSEM (Levy, 2011). The only two articles addressing model evaluation in BSEM are Hoofs et al. (2018), and Garnier-Villarreal and Jorgensen (2020), therefore the literature review was extended to the implementation of invariance test in BSEM to emphasis the model fit indices commonly utilized in applied research. Firstly, I describe the two main studies aforementioned, subsequently I extend this section to applied studies in different fields.

2.1 Simulation studies on fit measures

Hoofs et al. (2018) proposed a Bayesian variant of the root mean square error of approximation (RMSEA), they performed a simulation that manipulated several conditions such as misspecification, factor loadings magnitude, number of indicators , number of factors and sample size. Th results showed that 90% posterior probability interval of the BRMSEA is valid to evaluate model fit in large sample ($N \geq 1000$). In addition, the BRMSEA and RMSEA had a similar rejection pattern when sample size approach $N = 1000$, whereas models with larger misspecification and small sample size where not evaluated properly by the BRMSEA. However, the PPP showed to be adequate to evaluate models in small samples, but as expected and similar to its homologous χ^2, larger sample sizes were related to high rates of model rejection even when the misspecification was small. In this scenario, Hoofs et al. highlight the use of the BRMSEA to complement model evaluation along with

the PPP, given that BRMSEA is less sensitive when sample size increase compared to the PPP.

Garnier-Villarreal and Jorgensen (2020) tested a Bayesian formulation of TLI, CFI and $\hat{\gamma}$, they conducted a simultaion study based on the work by Hoofs et al. (2018) but they extended their work to different fit measures. In addition, Garnier-Villarreal and Jorgensen (2020) proposed a different Bayesian formulation to make BTLI, BCFI and B$\hat{\Gamma}$ closer to their frequentist version. Thus, MLE fit indices and Bayesian counterpart were compared showing that the sampling distributions of the posterior means of BTLI, BCFI and B$\hat{\gamma}$ are similar to their frequentist counterparts across sample sizes, model types, and levels of misspecification. The Bayesian fit indices allow overall model-fit evaluation using familiar metrics (MLE metrics). However, this study only includes non-informative priors given that BCFI, BTLI and B$\hat{\Gamma}$ require the computation of a null model. The null model in Bayesian SEM needs further discussion to understand its role and its meaning from a Bayesian perspective in the case of a null model with informative priors (Hoofs et al., 2018).

The most cited studies in invariance testing in SEM are the simulations performed by Cheung and Rensvold (2002), and Meade et al. (2008). The first article addressed for the first time optimal values to make decisions on keeping or rejecting the assumption of invariance between groups based not only on the χ^2 test, they evaluated 20 fit measures and recommended cutoff values to assess multigroup invariance, their main results showed that a ΔCFI smaller than or equal to 0.01 indicates that the null hypothesis of invariance should not be rejected. The second study by Meade et al. (2008) extends this work to add the power analysis of

several fit measures to detect the lack of invariance, as one of the main results they found out that a ΔCFI equals or larger than 0.002 is enough to reject the null hypothesis of invariance. They also provide a table with cutoff values for different fit measures that aims to help to create guidelines of best practices for conducting invariance tests in frequentist SEM.

2.2 Applied research

Van De Schoot et al. (2013) revisited the approximate measurement invariance approach suggested by Muthén and Asparouhov (2013) as a possible alternative to partial invariance, by allowing some "wiggle" room for the difference between groups, this flexible space is determined by the precision of the prior. Thus, an informative prior on the difference between groups is specified in order to allow some space around zero, without assuming a difference strictly equal to zero between groups.

Van De Schoot et al. (2013) tested the advantages of approximate MI in a real sample of psychologist and psychiatrist, regarding a new policy in Dutch mental health care. They showed that the latent means were different comparing approximate MI versus constrained maximum likelihood or a partial invariance model, therefore the conclusions about the difference between latent means would be different. Given these results, they also conducted a simulation to investigate the possible bias in the comparison of latent means when applying approximate MI. The results showed that approximate MI outperforms models where we assume full MI, especially when there exist many differences in the population on many intercepts. Approximate MI also showed less bias on the estimation of the latent means compared to partial invariance. It is important to clarify that Van De Schoot

et al. focused the study on the comparison of latent means between groups, therefore special importance was placed on intercept invariance assumption.

De Bondt and Van Petegem (2015) conducted an evaluation of the psychometric properties of the Overexcitability Questionnaire-Two (OEQ-II), in this study the authors implemented a BSEM analysis and tested the invariance measurement assumption utilizing approximate MI. The model fit assessment was performed evaluating the Posterior Predictive P-value (PPP) given that MPLUS and current software do bot provide alternative fit measures. Similarly, Bujacz et al. (2014) carry out an illustration on invariance testing in BSEM on the Hedonic and Eudaimonic Motives for Activities (HEMA) scale, the nested models were also evaluated based on the PPP and DIC, given the lack of literature on model fit indices in BSEM. However, when comparing the model fit of BSEM versus ML estimation, the PPP shows that the model does not fit the data while ML favors the same model and shows an appropriate fit. This result is hard to compare given the ML fit indices are not formulated to be close to the Bayesian PPP.

In addition, Cieciuch et al. (2018) implemented the approximate invariance approach to evaluate the Portrait Value Questionnaire (PVQ-21) included in the European Social Survey. Cieciuch et al. inlcuded 15 countries in the analysis, the nested models were evaluated based on the PPP value and its 95% credible interval.

CHAPTER 3

METHODS

A simulation study was performed to evaluate the BCFI, BRMSEA, BTLI and $B\hat{\Gamma}$, in addition several Bayesian fit indices will also be compared under different simulated conditions. The aim is to evaluate the new BCFI, BTLI and $B\hat{\Gamma}$ along with well known Bayesian fit measures in the context of measurement invariance between groups.

3.1 Simulation design

I simulated a four-factor two group CFA model with six indicators per factor. In each of the two population models, factors were standard normal ($\mu = 0$, $\sigma = 1$) with factor loadings of 0.70, all indicator intercepts were zero and residual variances were 0.51, hence indicators had unit variance. Factor 1 was simulated to be partially invariant. In the first condition Factor 1 had one non-invariant item, in the second condition Factor 1 comprised 3 non-invariant items, I included a large correlation between Factor 1 and factor 4 ($r = 0.8$), and a small correlation between Factor 1 and Factor 2 ($r = 0.2$). The rest of correlation between latent factors were set to $r = 0.5$ (see Figure 1 and Figure 2).

I simulated two additional conditions to evaluate the lack of weak invariance, and the lack of strong invariance. In the weak invariance condition a medium difference between groups in loadings was set to 0.26 based on the work by Meade et al. (2008). Secondly, in the lack of strong invariance condition a medium difference (0.4) in intercepts was added to the model based on Meade et al. (2008). Finally, sample size was manipulated to simulate a condition with a small number of

observation per group ($n = 100$), and a second condition with a large number observations per group ($n = 500$).

The weak and strong invariance condition where simulated in separate population models, when the lack of invariance was simulated the population model held the strong invariance assumption (equal intercepts between groups), similar condition was implemented in the lack of strong invariance, where the weak invariance assumption was held (equal factor loadings), but I manipulated the intercepts of the model, this was done based on the work by Cheung and Rensvold (2002) and Meade et al. (2008). Figure 1 depicts the population model for the condition of lack of weak invariance, all covariances between latent factors were estimated but they were not illustrated in the figure for the sake of clarity, omitted covariances were all set to ($r = 0.5$). Figure 2 illustrates the population models for the condition of lack of strong invariance, triangles represent the intercept with their parameter value. All covariances between latent factors were estimated but not included in the figure, omitted covariances were simulated with a parameter set to $r = 0.05$.

The function `simulateData` from the `lavaan` package (Rosseel, 2012) in `R` (R Core Team, 2018) was utilized to generate the data. The full factorial design includes 2 (sample size = 100/500) $\times$ 2 item condition (1 non-invariant/ 3 non-invariant) $\times$ 2 (weak invariance/ strong invariance) = 8 conditions. Models were fitted using MCMC estimation (Metropolis-Hastings) available in the `R` package `blavaan` (Merkle & Rosseel, 2018) via interface to Stan software (Carpenter et al., 2017) provided by `rstan` package (Stan Development Team, 2019).

I utilized the default non-informative priors set by Merkle and Rosseel (2018) to analyze the simulated data. The indicator intercepts were distributed as

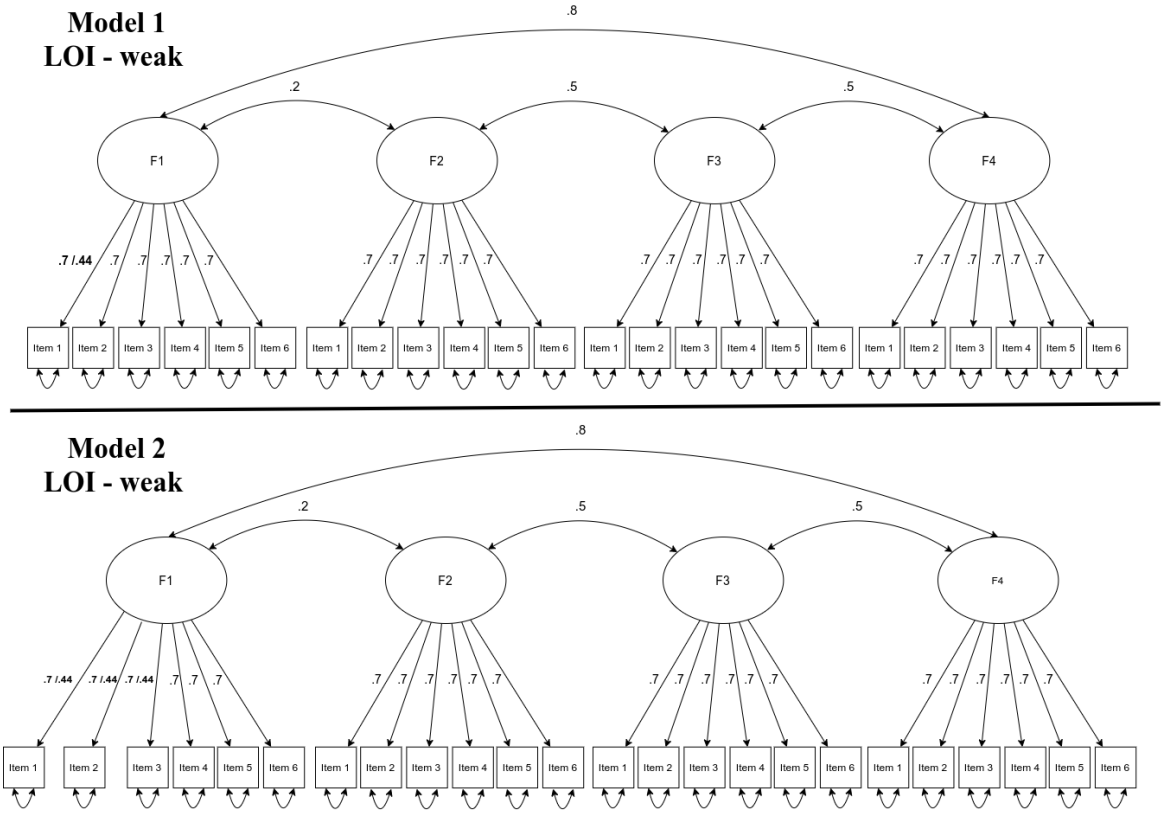

Figure 1: Population models for lack of weak invariance condition (LOI-weak)

$\sim N(\mu = 0, \sigma^2 = 100)$, indicator standard deviations were distributed as $\sim half - Cauchy(\mu = 0, \sigma^2 = 2.5)$.The latent variances and covariances were distributed as $\sim Inv - Wishart(\psi = I, \upsilon = nf + 1)$ where I represents an identity matrix of dimension equal to the number of factors(nf) , and degrees of freedom υ equal to the number of factors plus 1. Each model started with 10000 burned iterations, if the model did not converge based on the maximum R-hat (PSRF) value, a larger number of burned iterations was set until convergence was achieved.

The convergence of the Markov chains to the posterior distribution was evaluated utilizing the potential scale reduction factor (PSRF) (Gelman et al., 1996). Values close to 1.05 for each parameter were considered evidence of model convergence

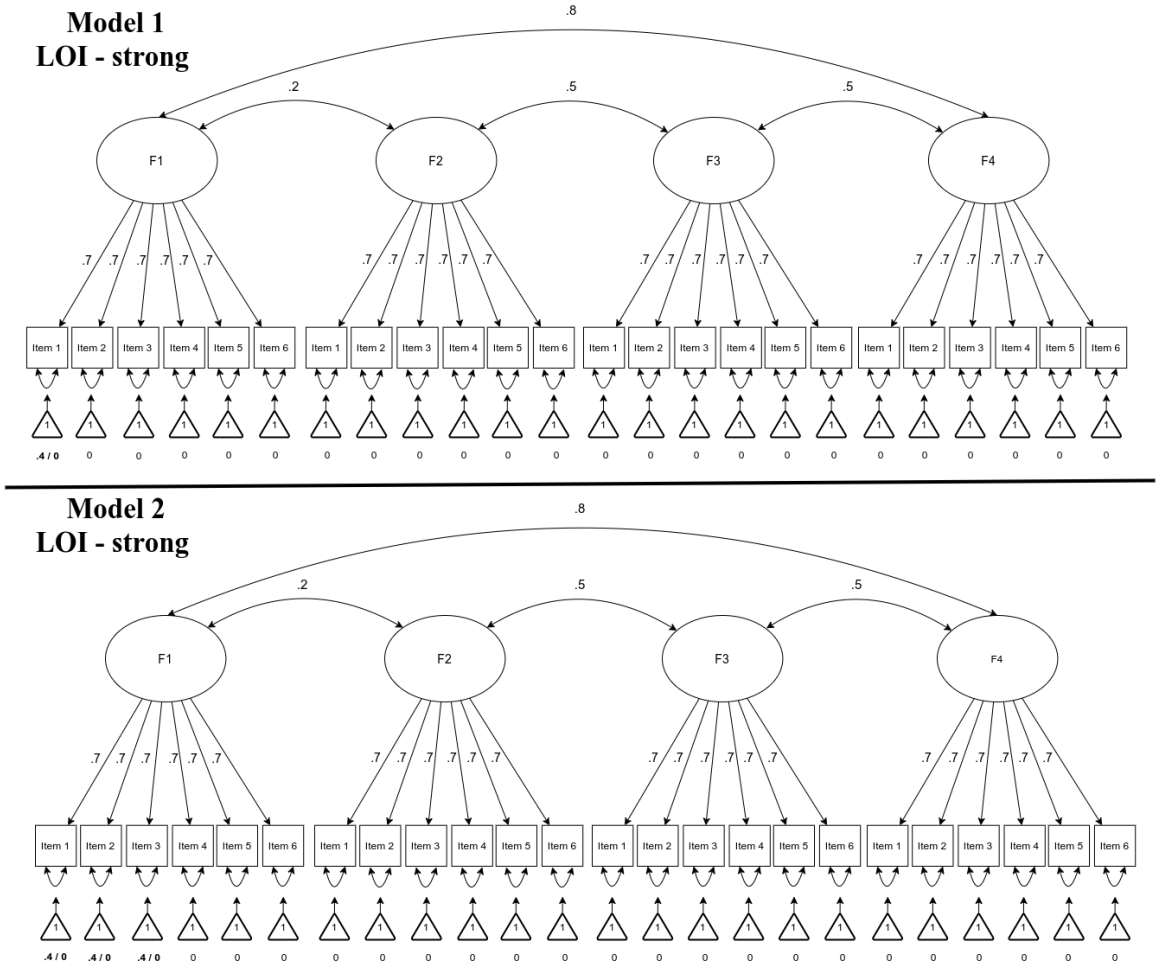

Figure 2: Population models for lack of strong invariance condition (LOI-strong)

(Gelman et al., 2013), 1000 samples were kept after burn-in for each chain, 3 chains were used for each estimated model. Thinning value was not included in the estimation.

The estimation of BTLI, BNFI and BCFI comprises the null model, thus the null model included was specified as the estimation of means and variances constraint to be equal between groups, following the recommendations by Little (2013) and Widaman and Thompson (2003).

Table 1

Simulated conditions and levels

Condition	Levels
Sample size per group	$n = 100$ or $n = 500$
Lack of invariance (LOI)	weak invariance (LOI-weak) or strong invariance (LOI-strong)
Number of manipulated items	3 non-invariant items or 1 non-invariant item

3.2 Analysis of results

Models were estimated using maximum likelihood estimation (MLE) and Bayesian estimation, correlations and effect size Cohen's d (Cohen, 1988) were reported to quantify the mean difference between MLE estimation and Bayesian estimation. In the case of the PPP, the power to achieve alpha level was reported following the recommendations by Muthén and Asparouhov (2012), and the work by Jorgensen et al. (2017). The mean credible intervals were also estimated to report mean differences between conditions.

The goodness of fit difference (ΔGFI) was estimated from the posterior distribution of each fit measure, I subtracted the fit index of model 1 (e.g. configural model) from the model 2 (e.g. weak model), hence: $\Delta GFI = Mod_2 - Mod_1$.

I also estimated the *probability of direction* which quantifies how probable is that the parameter is negative or positive, this estimation is performed with the posterior distribution values, and it is better define as the proportion of the posterior distribution that is of the median's sign (Makowski et al., 2019).

Finally, I performed a factorial ANOVA (2×2) to quantify the effect size (η^2) of each factor (number of manipulated items or sample size) on each fit measure and its ΔGFI.

CHAPTER 4

RESULTS

For each of the 8 conditions, I had 500 replications that converged for both estimations MLE and Bayesian models (PSRF < 1.05). Similar to Garnier-Villarreal and Jorgensen (2020), and following the recommendation by Vehtari et al. (2017) to estimate BSEM fit indices with $*p - pD$ in place of df, I selected to estimate pD_{Loo} instead of pD_{WAIC} or pD_{DIC}. Garnier-Villarreal and Jorgensen (2020) showed that estimated $*p - pD$ using noninformative piors should be close to the MLE degrees of freedom.

Table 2 shows the overall mean, effect size and correlation for the RMSEA, CFI, TLI, $\hat{\Gamma}$ and NFI. Means are reported for both ML estimation and Bayesian estimation. In both estimation methods, the standard deviation is similar suggesting a similar dispersion for both distributions (MLE and Bayesian), means are also close however, Bayesian fit indices tend to show a slightly worse fit. The Person correlation show a high relationship between MLE fit indices and the Bayesian indices, all correlations are higher than 0.90. Cohen's d was estimated following the formulation in Faul et al. (2007)[1], which shows a large effect according to Cohen's convention (Cohen, 1988). It is noteworthy, however, that the Cohen's d showed a considerably large effect because the standard deviations under MLE estimation and Bayesian inference are alike, therefore the effect sizes are remarkably large.

[1] $d_z = \frac{|\mu_z|}{\sigma_z} = \frac{|\mu_x - \mu_y|}{\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y}}$

Table 2

Comparison of Mean (SD) Fit Indices Using Maximum Likelihood and Bayesian Estimation

Index	MLE (SD)	Bayesian (SD)	ΔMean (SD)	Cohen's d	r
RMSEA	0.025 (0.014)	0.026 (0.014)	0.0003 (0.0007)	0.403	0.999
CFI	0.978 (0.019)	0.977 (0.020)	0.0010 (0.0009)	1.031	0.999
TLI	0.975 (0.023)	0.974 (0.023)	0.0009 (0.0010)	0.991	0.999
$\hat{\Gamma}$	0.983 (0.016)	0.982 (0.016)	0.0008 (0.0010)	0.805	0.998
$\hat{\Gamma}_{adj}$	0.980 (0.018)	0.978 (0.020)	0.0025 (0.0020)	1.229	0.998
NFI	0.854 (0.089)	0.856 (0.087)	0.0014 (0.0014)	1.004	1.000

Note: MLE = Maximum Likelihood Estimation; Bayesian = fit indices under Bayesian estimation; ΔMean = mean difference; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index; $\hat{\Gamma}$ = Gamma Hat; $\hat{\Gamma}_{adj}$ = Adjusted Gamma Hat; NFI = Normed Fit Index; r = Pearson correlation.

For the analysis of PPP, Muthén and Asparouhov (2012) suggested the possibility to use cut-off points similar to the frequentist alpha level (e.g. $< .10$, $< .05$ and $< .01$). Table 3 shows the proportion of models that met the alpha level respectively. The column $\%PPP$ reports the percentage of models that had a PPP value that meets the alpha level for the condition of lack of weak invariance and the lack of strong invariance. The results showed that there is larger power in the LOI-weak condition compare to LOI-strong condition, also statistical power increases when sample size is large ($N = 500$). However, in general the PPP does not achieve at least 80% of power on any of the manipulated conditions. This result matched previous simulations where PPP's statistical power rarely exceeded 80% unless a large sample is utilized, or strongly informative priors are added to the model (Jorgensen et al., 2017). The simulation that I designed comprises a maximum of 500 observations whereas Jorgensen et al. (2017) added an extra condition of 1000 observations, in this condition they were able to get a PPP's

statistical power close to 80%.

Table 3

Rejection rates by sample, and number of manipulated items

Manipulated items	Sample (N)	%PPP		Alpha level
		LOI - Weak	LOI-Strong	
1 item	100	10.00%	6.00%	$< .10$
	500	18.00%	6.00%	$< .10$
3 items	100	8.00%	5.00%	$< .10$
	500	36.00%	5.00%	$< .10$
1 item	100	4.00%	4.00%	$< .05$
	500	10.00%	3.00%	$< .05$
3 items	100	5.00%	3.00%	$< .05$
	500	21.00%	2.00%	$< .05$
1 item	100	1.00%	1.00%	$< .01$
	500	1.00%	0.00%	$< .01$
3 items	100	1.00%	0.00%	$< .01$
	500	4.00%	0.00%	$< .01$

Note: LOI = Lack of invariance

Regarding the mean values per simulated condition both LOI-Weak and LOI-Strong showed a similar pattern in their 90% credible intervals, across conditions the credible intervals are narrower when sample size increases. However, there are fit indices that are more sensitive to sample size such as Mc (see Table 4) which mean value is substantially larger when sample size is $N = 500$, similar result can be seen in Table 4 when comparing values for NFI. In addition PPP values were lower for the LOI-Weak model compare to LOI-Strong.

The mean difference of approximate fit indices was also estimated comparing two nested models. I calculated the difference between the model with LOI-Weak and its corresponded configural model, I repeated the same procedure to estimate the difference between LOI-Strong model and its corresponded weak invariance model,

Table 4

PPP and Bayesian Approximate Fit Indices by sample size and number of manipulated items

Index	Manipulated items	N	Mean [90% CI]	
			LOI-Weak	LOI-Strong
PPP	1 item	100	0.48 [0.05, 0.83]	0.418 [0.035, 0.790]
		500	0.34 [0.02, 0.72]	0.120 [0.001, 0.336]
	3 item	100	0.43 [0.06, 0.79]	0.349 [0.033, 0.726]
		500	0.21 [0.01, 0.53]	0.024 [0.000, 0.061]
BRMSEA	1 item	100	0.034 [0.014, 0.047]	0.036 [0.016, 0.048]
		500	0.010 [0.000, 0.018]	0.016 [0.007, 0.021]
	3 item	100	0.036 [0.017, 0.048]	0.038 [0.021, 0.050]
		500	0.013 [0.002, 0.020]	0.021 [0.014, 0.025]
$\hat{\Gamma}$	1 item	100	0.973 [0.947, 0.991]	0.970 [0.944, 0.988]
		500	0.997 [0.992, 1.000]	0.994 [0.989, 0.998]
	3 item	100	0.971 [0.947, 0.988]	0.966 [0.944, 0.985]
		500	0.995 [0.991, 0.999]	0.990 [0.984, 0.994]
$\hat{\Gamma_{adj}}$	1 item	100	0.966 [0.933, 0.989]	0.963 [0.932, 0.985]
		500	0.996 [0.990, 1.000]	0.992 [0.986, 0.997]
	3 item	100	0.963 [0.934, 0.985]	0.959 [0.932, 0.981]
		500	0.994 [0.988, 0.999]	0.988 [0.981, 0.993]
Mc	1 item	100	0.849 [0.715, 0.946]	0.831 [0.703, 0.930]
		500	0.982 [0.955, 0.999]	0.963 [0.934, 0.986]
	3 item	100	0.836 [0.717, 0.931]	0.812 [0.701, 0.912]
		500	0.973 [0.945, 0.994]	0.941 [0.910, 0.964]
CFI	1 item	100	0.966 [0.936, 0.988]	0.963 [0.932, 0.985]
		500	0.996 [0.990, 1.000]	0.992 [0.986, 0.997]
	3 item	100	0.961 [0.932, 0.984]	0.958 [0.929, 0.981]
		500	0.994 [0.988, 0.999]	0.988 [0.981, 0.992]

Note: PPP = posterior predictive p-value; BRMSEA = Bayesian Root Mean Square Error of Approximation; $\hat{\Gamma}$ = Gamma hat; $\hat{\Gamma}_a dj$ = Adjusted gamma hat; Mc = McDonald's centrality index, CFI = comparative fit index, TLI = Tucker-Lewis Index; NFI = normed fit index.

Table 4. Continued

Index	Manipulated items	N	Mean [90% CI]	
			LOI-Weak	LOI-Strong
TLI	1 item	100	0.961 [0.925, 0.986]	0.958 [0.924, 0.983]
		500	0.996 [0.988, 1.002]	0.991 [0.984, 0.997]
	3 item	100	0.956 [0.921, 0.981]	0.953 [0.920, 0.979]
		500	0.993 [0.986, 0.999]	0.986 [0.979, 0.991]
NFI	1 item	100	0.772 [0.743, 0.794]	0.767 [0.739, 0.787]
		500	0.946 [0.940, 0.951]	0.942 [0.935, 0.946]
	3 item	100	0.764 [0.735, 0.785]	0.763 [0.735, 0.784]
		500	0.942 [0.936, 0.947]	0.937 [0.931, 0.943]

Note: PPP = posterior predictive p-value; BRMSEA = Bayesian Root Mean Square Error of Approximation; $\hat{\Gamma}$ = Gamma hat; $\hat{\Gamma}_a dj$ = Adjusted gamma hat; Mc = McDonald's centrality index, CFI = comparative fit index, TLI = Tucker-Lewis Index; NFI = normed fit index.

the final step calculated the mean of those difference across iterations along with the 95% credible interval. Table 5 reports the results of comparing LOI-Weak and LOI-Strong models versus configural invariance, and weak invariance model respectively. The BRMSEA mean difference when sample size is small ($N = 100$) does not show a large difference between models, it is noteworthy that under the conditions of 1 manipulated item and $N = 100$ the mean difference was negative ($M = -.000[-0.016, -0.015]$) when detecting lack of weak invariance, this result reveals that BRMSEA was not sensitive to detect the misfit under the simulated conditions. However, the difference is larger when sample size is larger ($N = 500$) showing that BRMSEA has larger power to detect a misfit when sample size increses.

BRMSEA shows relevant differences when contrasting credible intervals between models. The credible intervals in the case of the LOI-Strong shows only positive

values when $N = 500$ whereas the credible intervals in the LOI-Weak model comprises only negative values (Table 5). This pattern denotes that BRMSEA accurately detected a misfit in the model under the strong invariance test with a large sample size.

Gamma Hat index ($\hat{\Gamma}$) showed a small mean difference in the LOI-weak model compare to LOI-strong model when only 1 item was misfitted and sample size was small, credible intervals are wider under small sample size condition. Similarly, the adjusted index ($\hat{\Gamma}_{adj}$) evidence a larger mean difference in the LOI-strong model compare to LOI-weak model but it is noticeable that credible intervals do not comprise zero when sample size is large on both the adjusted and non-adjusted measures.

The McDonald's centrality index (Mc) in Table 5 also shows smaller mean differences in the LOI-weak model compare to LOI-strong, however the credible intervals in the LOI-weak contain zero even when sample size is large, this characteristic is not present in LOI-strong where credible intervals do not cover zero value when sample size is large. Bayesian CFI (BCFI) also has a similar result to Mc.

In Table 5 the BTLI had a positive mean ($M = .001$, $[-0.028, 0.031]$) when sample size is small and only 1 item is misfitted between groups, similar to the case of the BRMSEA; BTLI shows that was not sensitive to the simulated magnitude of misfit when sample size was small or the model did not have more sources of misfit. For instance, the condition where sample is small but number of manipulated items were three misfitted items ($M = -0.002$, $[-0.032, 0.029]$), even though the number of misfitted parameters was three, BTLI did not show enough sensitivity to detect

the misfit.

The Bayesian normed fit index (BNFI) showed a large difference in the LOI-strong model compare to the LOI-weak model. Large number of observations helped the BNFI to accurately capture the presence of misfit in the model in the LOI-weak model given that credible intervals are narrowed and exclude zero. But in the LOI-strong model, BNFI was able to capture the the existence of misfit regardless of sample size (see Table 5).

Table 5

Invariance Mean Differences of Bayesian Approximate Fit Measures

Index	Manipulated items	N	ΔMean [95% CI]	
			LOI-Weak	LOI-Strong
BRMSEA	1 item	100	-0.000 [-0.016, 0.015]	0.003 [-0.011, 0.016]
		500	0.003 [-0.005, 0.012]	0.009 [0.002, 0.016]
	3 item	100	0.001 [-0.014, 0.016]	0.006 [-0.007, 0.019]
		500	0.006 [-0.001, 0.014]	0.015 [0.008, 0.021]
$B\hat{\Gamma}$	1 item	100	-0.000 [-0.019, 0.018]	-0.004 [-0.021, 0.013]
		500	-0.001 [-0.004, 0.002]	-0.004 [-0.008, -0.001]
	3 item	100	-0.002 [-0.021, 0.017]	-0.009 [-0.026, 0.009]
		500	-0.003 [-0.006, 0.000]	-0.008 [-0.012, -0.005]
$B\hat{\Gamma}_{adj}$	1 item	100	0.001 [-0.023, 0.025]	-0.004 [-0.025, 0.017]
		500	-0.001 [-0.005, 0.002]	-0.005 [-0.009, -0.002]
	3 item	100	-0.001 [-0.025, 0.023]	-0.009 [-0.030, 0.012]
		500	-0.003 [-0.008, 0.001]	-0.010 [-0.014, -0.006]
BMc_{adj}	1 item	100	-0.002 [-0.101, 0.098]	-0.023 [-0.114, 0.069]
		500	-0.007 [-0.026, 0.010]	-0.026 [-0.045, -0.008]
	3 item	100	-0.011 [-0.110, 0.088]	-0.045 [-0.135, 0.046]
		500	-0.016 [-0.036, 0.003]	-0.050 [-0.068, -0.031]
BCFI	1 item	100	-0.001 [-0.025, 0.024]	-0.005 [-0.027, 0.016]
		500	-0.002 [-0.006, 0.002]	-0.006 [-0.009, -0.002]
	3 item	100	-0.003 [-0.028, 0.023]	-0.011 [-0.032, 0.011]
		500	-0.004 [-0.008, 0.001]	-0.010 [-0.014, -0.007]

Note: BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_adj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; 90% CI = the average lower and upper bounds; LOI = lack of invariance.

Table 5 Continued

Index	Manipulated items	N	ΔMean [95% CI]	
			LOI-Weak	LOI-Strong
BTLI	1 item	100	0.001 [-0.028, 0.031]	-0.005 [-0.030, 0.020]
		500	-0.002 [-0.008, 0.004]	-0.007 [-0.012, -0.002]
	3 item	100	-0.002 [-0.032, 0.029]	-0.011 [-0.036, 0.014]
		500	-0.005 [-0.011, 0.001]	-0.013 [-0.018, -0.008]
BNFI	1 item	100	-0.008 [-0.027, 0.011]	-0.011 [-0.028, -0.028]
		500	-0.004 [-0.009, 0.001]	-0.008 [-0.012, -0.012]
	3 item	100	-0.010 [-0.030, 0.009]	-0.015 [-0.032, -0.032]
		500	-0.006 [-0.011, -0.001]	-0.013 [-0.017, -0.017]

Note: BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_a dj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; 90% CI = the average lower and upper bounds; LOI = lack of invariance.

Table 6 shows the probability of direction (PD) for the approximate fit measures by condition, in the case of the estimation that I performed, it quantifies the proportion of mean differences above 0, for instance when calculating the difference (e.g. ΔBRMSEA) between two models such as weak invariance versus strong invariance model, I estimated the proportion of times the difference between weak model and configural model is larger than zero. In the case of BRMSEA the ideal values would be large proportions denoting that weak model has a large BRMSEA value compare to the BRMSEA value of the configural model. Nonetheless, all measures but BRMSEA are expected to be close to 0 in order to claim that most of the times the fit measure detects any source of misfit. Thus, similar to the reported findings in Table 5, BRMSEA is not sensitive to detect misfit between groups when sample size is small especially at testing the lack of weak invariance. However, BRMSEA showed a large proportion of values above 0 in the LOI-strong model, the

smallest mean proportion was $M = 0.635, SD = 0.147$.

The measures $\hat{\Gamma}$ and $\hat{\Gamma}_{adj}$ showed more power to detect a misfit when sample size is large, equivalently BMc, BCFI, BTLI and BNFI had a similar tendency. But is remarkable that BNFI had more sensitivity to detect the presence of misfit when 1 item was manipulated in a small sample condition, by looking at the mean proportion, in the LOI-weak model the value was M = 0.208 (SD= 0.080) and for the LOI-strong was M = 0.116 (SD = 0.074).

Table 6

Mean of Probability of direction (PD) of LOI-weak and LOI-strong models by simulated conditions

Index	Manipulated items	N	M (SD)	
			LOI-weak	LOI-strong
BRMSEA	1 item	100	0.469 (0.123)	0.635 (0.147)
		500	0.644 (0.246)	0.959 (0.126)
	3 item	100	0.531 (0.135)	0.772 (0.146)
		500	0.849 (0.201)	0.999 (0.017)
$B\hat{\Gamma}$	1 item	100	0.476 (0.121)	0.318(0.138)
		500	0.180 (0.122)	0.011(0.024)
	3 item	100	0.415 (0.125)	0.188(0.118)
		500	0.068 (0.071)	0.000(0.000)
$B\hat{\Gamma}_{adj}$	1 item	100	0.517 (0.123)	0.355(0.143)
		500	0.194 (0.129)	0.013(0.026)
	3 item	100	0.457 (0.130)	0.217 (0.127)
		500	0.076 (0.077)	0.000 (0.000)
BMc	1 item	100	0.476 (0.121)	0.318 (0.138)
		500	0.180 (0.122)	0.011 (0.024)
	3 item	100	0.415 (0.125)	0.188 (0.118)
		500	0.068 (0.071)	0.000 (0.000)
BCFI	1 item	100	0.476 (0.121)	0.318 (0.138)
		500	0.180 (0.122)	0.011 (0.024)
	3 item	100	0.415 (0.125)	0.188 (0.118)
		500	0.068 (0.071)	0.000 (0.000)

Note: BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_adj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; LOI-weak = lack of weak invariance; LOI-strong = lack of strong invariance.

Table 6. Continued

Index	Manipulated items	N	M (SD)	
			LOI-weak	LOI-strong
BTLI	1 item	100	0.523 (0.116)	0.357 (0.142)
		500	0.238 (0.121)	0.014 (0.029)
	3 item	100	0.462 (0.127)	0.220 (0.128)
		500	0.088 (0.083)	0.000 (0.000)
BNFI	1 item	100	0.208 (0.080)	0.116 (0.074)
		500	0.062 (0.048)	0.002 (0.006)
	3 item	100	0.168 (0.074)	0.057 (0.053)
		500	0.016 (0.023)	0.000 (0.000)

Note: BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_a dj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; LOI-weak = lack of weak invariance; LOI-strong = lack of strong invariance.

Table 7 shows the effect sizes for each model (LOI-weak and LOI-strong). Based on Cohen (1988) the criteria to interpret the size of η^2 is: negligible $< 0.01 <$ small $< 0.06 <$ moderate $< 0.14 <$ large. It is noticeable that N has a large effect on the approximate fit measures the effect sizes ranged from $\eta^2 = 0.585$ to $\eta^2 = 0.979$ in the LOI-weak model, and $\eta^2 = 0.404$ to $\eta^2 = 0.979$. The PPP was not affected by number of manipulated items but there was a large effect of sample size when analyzing the LOI-strong model ($\eta^2 = 0.374$) and a moderate effect in the LOI-weak model($\eta^2 = 0.094$). The ΔBRMSEA was considerably affected by sample size ($\eta^2 = 0.363$) and moderately affected by the number of manipulated items ($\eta^2 = 0.092$) in the LOI-weak model, similar tendency is observed in LOI-strong model, but in this case the impact of number of manipulated items ($\eta^2 = 0.137$) and sample size ($\eta^2 = 0.494$) is larger than LOI-weak model (see Table 7).

Finally, Table 8 reports the effect sizes (η^2) of each condition on the probability of

Table 7

Proportions of variance (η^2) explained by simulated conditions

Fit Measure	N items	N	$N \times items$	N items	N	$N \times items$
	LOI-weak			LOI-strong		
ΔBRMSEA	**0.092**	**0.363**	0.021	**0.137**	**0.494**	0.012
$\Delta B\hat{\Gamma}$	**0.078**	0.018	0.000	**0.285**	0.000	0.000
$\Delta\hat{\Gamma}_{adj}$	**0.072**	**0.102**	0.000	**0.284**	0.015	0.000
ΔBMc_{adj}	**0.082**	0.029	0.000	**0.297**	0.010	0.000
ΔBCFI	**0.080**	0.016	0.000	**0.281**	0.000	0.000
ΔBTLI	**0.078**	**0.145**	0.000	**0.290**	0.031	0.000
ΔBNFI	**0.079**	**0.366**	0.001	**0.283**	**0.140**	0.002
PPP	0.018	**0.094**	0.015	0.026	**0.374**	0.001
Bayesian Factor	**0.181**	**0.220**	**0.075**	**0.235**	**0.404**	**0.101**
BRMSEA	0.008	**0.642**	0.001	0.024	**0.573**	0.003
Bayesian $\hat{\Gamma}$	0.004	**0.601**	0.000	0.000	**0.617**	0.000
Bayesian $\hat{\Gamma}_{adj}$	0.004	**0.601**	0.000	0.014	**0.569**	0.000
Bayesian Mc	0.004	**0.619**	0.000	0.015	**0.583**	0.000
Bayesian CFI	0.006	**0.600**	0.001	0.015	**0.568**	0.000
Bayesian TLI	0.007	**0.585**	0.001	0.015	**0.557**	0.000
Bayesian NFI	0.001	**0.979**	0.000	0.001	**0.980**	0.000

Note: N of items = number of manipulated items; $N \times items$ = interaction term between sample size and N of Items; N = sample size; Δ = difference; BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_adj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; LOI-weak = lack of weak invariance; LOI-strong = lack of strong invariance; Effects larger than %6 of variance explained are in bold.

direction. Mainly, each PD is largely affected by sample size, however only BTLI and BRMSEA were moderately affected by number of items in the LOI-weak model.

Table 8

Proportions of variance (η^2) explained of probability of direction (PD) by simulated conditions

Fit Measure	N of Items	N	$N \times items$	N of Items	N	$N \times items$
	LOI-weak			LOI-strong		
PD BRMSEA	**0.093**	**0.275**	0.029	0.054	**0.524**	0.016
PD $B\hat{\Gamma}$	0.044	**0.632**	0.006	0.048	**0.593**	0.034
PD $B\hat{\Gamma}_{adj}$	0.040	**0.657**	0.006	0.045	**0.625**	0.031
PD BMc	0.044	**0.632**	0.006	0.048	**0.593**	0.034
PD BCFI	0.044	**0.632**	0.006	0.048	**0.593**	0.034
PD BTLI	**0.067**	**0.611**	0.014	0.046	**0.629**	0.030
PD BNFI	0.048	**0.554**	0.001	0.053	**0.424**	0.046

Note: N of items = number of manipulated items; $N \times items$ = interaction term between sample size and N of Items; N = sample size; BRMSEA = Bayesian Root Mean Square Error of Approximation; $B\hat{\Gamma}$ = Bayesian Gamma hat; $B\hat{\Gamma}_adj$ = Bayesian adjusted gamma hat; BMc = Bayesian McDonald's centrality index, BCFI = Bayesian comparative fit index, BTLI = Bayesian Tucker-Lewis Index; BNFI = Bayesian normed fit index; LOI-weak = lack of weak invariance; LOI-strong = lack of strong invariance; Effects larger than %6 of variance explained are in bold.

CHAPTER 5
DISCUSSION

The simulation described in this project is a parsimonious attempt to give some evidence of the behavior of new fit measures in a growing statistical theory such as Bayesian Structural Equation Modeling. Results should be interpreted as an approximation to evaluate the Bayesian fit measures as proposed by Garnier-Villarreal and Jorgensen (2020). In addition, the simulated conditions where not enough to evaluate several possible scenarios as larger number of observations, different levels of misfit or more misfitted items between groups. These are relevant additional conditions that should be explored in future applications or simulations contemplating approximate fit measures in BSEM. However, this first exploration intends to be the beginning of a larger project where the novel Bayesian fit measures are tested under different hypothesis and scenarios in the context of invariance testing.

I have found large correlations between the Bayesian approximate fit measures and its MLE counterpart (Table 2). Garnier-Villarreal and Jorgensen (2020) also found large correlations between MLE and Bayesian fit measures, however the correlations found in this study were slightly higher than the correlations in reported in Garnier-Villarreal and Jorgensen (2020), the lack of informative priors might be a cause for this small difference in studies. Also conditions are different in both studies, in my design I conducted a multigroup comparison whereas Garnier-Villarreal and Jorgensen (2020) tested all fit measures in a single group simulation. The large relationship between Bayesian and MLE fit measures under

non-informative priors in multigroup analysis allows to compare both approaches in future developments.

The PPP showed a remarkable lack of power to detect misspecification, similar result was found by Jorgensen et al. (2017). Their simulation contemplated more levels for the sample size condition and they also manipulated the prior's level of information. Their findings showed that PPP was able to diagnose misfit when sample size is large and misfit is severe. Giving that I simulated a medium size misfit is likely that PPP would have more power in a condition with a larger magnitude of misfit. In addition, the effect of sample size on PPP (Table 7) denotes that it is moderately to largely influenced by number of observations.

Regarding the goodness of fit difference (ΔGFI) in Table 5 differences are large compare to the study by Cheung and Rensvold (2002) but in this study I included a group difference (misfit) in loadings and intercepts which helps to diagnose the power to detect the source of misfit. Cheung and Rensvold (2002) recommended, for example, to use critical values $\Delta\hat{\Gamma} \leq .001$, and ΔCFI $\leq .01$ as possible values to retain the null hypothesis of invariance. Meade et al. (2008) based on the work by Cheung and Rensvold (2002) proposed stricter critical values $\Delta\hat{\Gamma} \leq .001$, and ΔCFI $\leq .002$. However, more conditions must be added to evaluate the evidence that might support the critical values mentioned by Meade et al. (2008). In addition, Cheung and Rensvold (2002) only evaluated Type I error whereas my design focused on statistical power (Type II error). These differences between studies makes difficult to support the application of previous cut-off points to BSEM. Type II error should also be evaluated in future studies related to approximate fit indices in BSEM.

The medium size difference in the intercepts was more sensitive to be detected than the misfit on the loadings. When looking at the credible intervals (Table 5) all ΔGFI performed better to detect misfit in the strong invariance test when $N = 500$ as credible intervals did not cover zero, regardless of number of manipulated items. However, the ANOVA analysis showed that there are ΔGFI more affected by sample size (7): such as ΔBRMSEA, $\Delta B\hat{\Gamma}_{adj}$, ΔBTLI and ΔBNFI. Meade et al. (2008) also found an effect of sample size on ΔBRMSEA but it was a small effect $\omega^2 = 0.01$. In the present study I've found a moderate to large effect of sample size on the aforementioned indices, this is evidence that $\Delta B\hat{\Gamma}$ and BCFI could be better candidates to evaluate the assumption of invariance giving that they are less impacted by sample size. The CFI in MLE has especially shown to be fairly independent of sample size (Fan & Sivo, 2007; Hu & Bentler, 1998).

The interaction number of manipulated and sample size did not largely influenced any ΔGFI, result that matches Meade et al. (2008) findings in MLE factor analysis. Surprisingly, the Bayesian factor was the only index impacted largely by the interaction of this two conditions. I decided to include the Bayesian factor although it is not a fit measure, it has been evaluated as possible test to make decisions in invariance testing (Verhagen & Fox, 2013; Verhagen et al., 2016). Bayesian factors are used to evaluate the evidence that we have to support a model versus another alternative model based on likelihood. In this case, the package `blavaan` (Merkle & Rosseel, 2018) includes the log-Bayes factor and it is moderately influenced by the interaction between number of items manipulated and sample size. This reveals that BF needs further evaluation in SEM models.

Preliminary guidelines for researchers based on the results might comprise the

implementation of B$\hat{\Gamma}$ and BCFI for conducting invariance testing in Bayesian inference. These two fit measures showed to be less influenced by sample size, and better estimates to detect misfit even under one misfitted item condition. It is important to be cautious about the acceptance of previous MLE guidelines when testing invariance in Bayesian SEM giving that the result showed that the GFI's behavior under frequentist approach can differ compare to Bayesian inference.

References

Barendse, M., Albers, C., Oort, F., & Timmerman, M. (2014). Measurement bias detection through bayesian factor analysis. *Frontiers in psychology*, *5*, 1087.

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

Brooks, S. P., & Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455. https://doi.org/10.1080/10618600.1998.10474787

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. Guilford Publications.

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bujacz, A., Vittersø, J., Huta, V., & Kaczmarek, L. D. (2014). Measuring hedonia and eudaimonia as motives for activities: Cross-national investigation through traditional and bayesian structural equation modeling. *Frontiers in Psychology*, *5*, 984.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). *STAN* : A Probabilistic Programming Language. *Journal of Statistical Software*, *76*(1). https://doi.org/10.18637/jss.v076.i01

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural equation modeling*, *9*(2), 233–255.

Chiorri, C., Day, T., & Malmberg, L.-E. (2014). An approximate measurement invariance approach to within-couple relationship quality. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00983

Cieciuch, J., Davidov, E., Algesheimer, R., & Schmidt, P. (2018). Testing for Approximate Measurement Invariance of Human Values in the European Social Survey. *Sociological Methods & Research*, *47*(4), 665–686. https://doi.org/10.1177/0049124117701478

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale.

De Bondt, N., & Van Petegem, P. (2015). Psychometric evaluation of the overexcitability questionnaire-two applying bayesian structural equation modeling (bsem) and multiple-group bsem-based alignment with approximate measurement invariance. *Frontiers in psychology*, *6*, 1963.

Fan, X., & Sivo, S. A. (2007). Sensitivity of Fit Indices to Model Misspecification and Model Types. *Multivariate Behavioral Research*, *42*(3), 509–529. https://doi.org/10.1080/00273170701382864

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Garnier-Villarreal, M., & Jorgensen, T. D. (2020). Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*, *25*(1), 46–70. https://doi.org/10.1037/met0000224

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733–760.

Hoofs, H., van de Schoot, R., Jansen, N. W. H., & Kant, I. (2018). Evaluating Model Fit in Bayesian Confirmatory Factor Analysis With Large Samples: Simulation Study Introducing the BRMSEA. *Educational and Psychological Measurement*, *78*(4), 537–568. https://doi.org/10.1177/0013164417709314

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

Hu, L.-t., & Bentler, P. M. (1998). Fit Indices in Covariance Structure Modeling: Sensitivity to Underparameterized Model Misspecification, 30.

Jorgensen, T. D., Garnier-Villarreal, M., Pornprasermanit, S., & Lee, J. (2017). Small-variance priors can prevent detecting important misspecifications in bayesian confirmatory factor analysis, In *The annual meeting of the psychometric society*. Springer.

Kaplan, D. (2014). *Bayesian statistics for the social sciences*. New York, NY, US, Guilford Press.

Kline, R. B. (2015). *Principles and Practice of Structural Equation Modeling, Fourth Edition*. Guilford Publications.

Kruschke, J. K. (2010). Bayesian data analysis: Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. https://doi.org/10.1002/wcs.72

Lee, S.-Y. (2007). *Structural equation modeling: A bayesian approach* (Vol. 711). John Wiley & Sons.

Lee, S.-Y., & Song, X.-Y. (2012). *Basic and advanced Bayesian structural equation modeling: With applications in the medical and behavioral sciences*. John Wiley & Sons.

Levy, R. (2011). Bayesian Data-Model Fit Assessment for Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(4), 663–685. https://doi.org/10.1080/10705511.2011.607723

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.

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford press.

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19–35. https://doi.org/10.1037/1082-989X.11.1.19

Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., & Lüdecke, D. (2019). Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02767

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*(1), 97–103. https://doi.org/10.1007/BF01908590

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of applied psychology*, *93*(3), 568.

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

Merkle, E. C. (2011). A Comparison of Imputation Methods for Bayesian Factor Analysis Models. *Journal of Educational and Behavioral Statistics*, *36*(2), 257–276. https://doi.org/10.3102/1076998610375833

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*(4), 1–30. https://doi.org/10.18637/jss.v085.i04

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

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

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*(4), 407–419.

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Robert, C., & Casella, G. (2010). *Introducing Monte Carlo Methods with R*. New York, Springer-Verlag.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman; Hall/CRC.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003

Song, X.-Y., & Lee, S.-Y. (2012). *Basic and Advanced Bayesian Structural Equation Modeling: With Applications in the Medical and Behavioral Sciences*. Wiley. https://doi.org/10.1002/9781118358887

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. V. D. (2002a). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002b). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Stan Development Team. (2019). RStan: The R interface to Stan [R package version 2.19.2]. http://mc-stan.org/

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1–10.

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*, 770.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Efficient implementation of leave-one-out cross-validation and waic for evaluating fitted bayesian models. *arXiv preprint arXiv:1507.04544*.

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. https://doi.org/10.1007/s11222-016-9696-4

Verhagen, A. J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *British Journal of Mathematical and Statistical Psychology*, *66*(3), 383–401. https://doi.org/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. https://doi.org/10.1016/j.jmp.2015.06.005

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods*, *8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16