February 2019

Exploring the Behavior of Model Fit Criteria in the Bayesian Approximate Measurement Invariance: A Simulation Study

Abeer Atallah S. Alamri
University of South Florida, alamri.research@gmail.com

Exploring the Behavior of Model Fit Criteria in the Bayesian Approximate Measurement Invariance: A Simulation Study

by

Abeer Atallah S. Alamri

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction with emphasis in
Measurement and Evaluation
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: Eun Sook Kim, Ph.D.
John Ferron, Ph.D.
Jennifer Wolgemuth, Ph.D.
Stephen Stark, Ph.D.

Date of Approval:
November 28,2018

Keywords: Bayesian, approximate measurement invariance, cross-cultural, prior, Bayes factor

Dedication

This dissertation is dedicated

To my parents **Attalah Saleh Alamri** and **Tala Oadah Alamri** who invested so much in me.

I express my special gratitude and deep appreciation dedicating this achievement. Thank you for making me a better person and helping me to feel proud of what I am doing.

إلى والديّ الحبيبين:

عطاالله صالح العمري وَ طلعة عودة العمري

شكرًا لكل اهتمام و تقدير و تحفيز وهبتوني إياه لأكون إمرأة فخورة بنفسها

أهدي لكما هذا الجهد والإنجاز و الأمل، راجية به أن أهبكما نصيبكما من الفخر

Acknowledgements

First and foremost, I would like to thank my *Lord*, **Allah**, whose favor on me is always great, for giving me the power and the knowledge to pursue my doctoral study abroad. Without His support and mercy, I would not have been able to make it through this program.

Along the Ph.D. journey, I have met so many people who nudged me in the right direction and encouraged me to continue. I would like to express my sincere appreciation to my major advisor, **Dr. Eun Sook Kim**, for her thorough understanding and endless support throughout the dissertation process and to my entire doctoral study. Her patience and continuous guidance made it a thoughtful and rewarding journey. She inspired me in methodological research and motivated me to participate in a variety of research activities. Her mentorship and encouragement were paramount in helping me broaden and sharpen my skills and obtain research experiences. She has guided me on how to conduct research and how to receive other academic opinions and schools of thought first with respect and then with critical thinking. I would give Dr. Kim most of the credit for becoming the kind of psychometrician I am today.

I would like to express my special thanks to my dissertation committee members, **Professor John Ferron, Dr. Jennifer Wolgemuth, and Professor Stephen Stark** for generously sharing their time to provide constructive comments and crucial remarks that shaped and improved the quality of my dissertation. I also wish to express my heartfelt thanks to **Professor Jeffrey Kromrey**, who taught me how to smoothly move statistics from classroom to my research in real-world. His impact stayed with me even after his retirement.

I want to thank **Professor Robert Dedrick** and **Dr. Sarah Kiefer** not only for sharing their broad knowledge and expertise but also for providing truly warm welcomes and considerable encouragements that meant a lot to an international female student. I feel fortunate to have been in such a caring and supportive research community.

Third and importantly, completing my dissertation would not have been possible without the love, patience, and support of my family in the United States and Saudi Arabia. Special

thanks go to my dear husband, **Dr. Abdullah Aljohani**, who has given me constant support and love during the completion of this journey. My debts to my soulmates **Eng. Bader, Eng. Bayan, and IT Specialist Juman** are too many to count. Thank you for being independent, influential, and flourishing into your own. I have said it always; you are a part of this success; thank you for your unconditional love and sacrifices. I am forever grateful to my future hope **Fahad and Khalid** for being healthy, happy, and understanding for the time away from them during my study. When I look at all they have done for me; I know that I am truly blessed.

I would like to express my indebtedness to **all my siblings**, especially my sisters **Laila and Maha**. You kept calling and visiting me in the United States., have supported me and trusted in me unconditionally. Sometimes it was your unwavering confidence that I can complete this degree that motivated me to continue, in the face of my own doubts. Thank you, my little princess **Lana**, and my nephew **Nawaf**, it is your magic that pulled me out of many intervals of depression. You made my last two years in the United States so special.

Special thanks to my fellow graduate student who has always been a major source of support when things would get a bit discouraging, **Dr. Yan Wang**. From coding advice to watching my practice presentations to being an invaluable coauthor. I really enjoyed collaborating with her in preparing conference presentations and journal publications.

Special appreciation goes to **Prince Dr. Faisal bin Abdullah Al-Mashary Al- Saud**, the Executive Officer of Qiyas and the President of the Education and Training Evaluation Commission, who has believed in me like nobody else, motivated me to persist and offered me his unwavering backing.

Last but not least, I am grateful to my beloved country **Kingdom of Saudi Arabia** and the **National Center for Assessment (Qiyas)** in Riyadh, Saudi Arabia, for supported me morally, academically and financially that I otherwise would not have been able to achieve my academic goals.

All of you, thank you, pursuing my Ph.D. degree was one of the most important and informative experiences in my life. I sincerely appreciate your support and encouragement throughout my graduate studies.

Table of Contents

List of Tables

List of Figures

Abstract

Measurement invariance (MI) is conducted to ensure that differences found in the results of group comparisons are due to true substantive differences and not methodological artifacts. Previous cross-cultural and cross-national studies with large number of groups showed that the advanced measurement invariance level was rarely held when utilizing the traditional (frequentist) MI approach. The Bayesian approximate measurement invariance (BAMI) was introduced to override the traditional MI strict assumption, because trivial non-invariance in parameters across groups is allowed. Although the concept of the BAMI, which has been utilized since 2013, was incorporated into the context of structural equation modeling, there is still a need for clear-cut criteria of BAMI for group comparison because the Bayesian approach can account for uncertainty when appropriately modeled.

Given this, the current study demonstrates the usefulness and flexibility of Bayesian approximate measurement invariance and aims to examine the extent to which employing different research settings would affect the behavior of the BAMI across populations. Particularly, a Monte Carlo study was designed to evaluate the sensitivity of the BAMI model fit criteria to varying prior estimates and simulation conditions. The design factors include the group numbers, percent of groups with the non-invariant item intercepts (balanced and unbalanced), and magnitude and directions of DIF item intercepts. The conditions were chosen based on a systematic literature review of the BAMI applied studies conducted between 2013 and 2017 as well as a review of the BAMI published simulation studies. Crossing all the data

generation factors for exact models resulted in a total of 2 simulation conditions, whereas approximate models resulted in a total of 24 simulation conditions. Primarily, the analysis procedure included two modeling approaches. a) exact-zero scalar MI against exact-zero metric MI, and b) Bayesian approximate-zero scalar MI with five level of prior precision variances. The generated data were analyzed using maximum likelihood estimator and Bayes estimator with five different prior variances that were addressed in the literature, .001, .005, .01, .05, and .10. All generated data were fitted to each model. Two BAMI model fit criteria were used (PPP and 95% CI) as well as three model comparisons criteria (Bayes factor, BIC, and DIC). In order to assess the sensitivity of the exact and BAMI model fit criteria, three outcome variables were evaluated as a function of design factors: (a) convergence rates, (b) model fit evaluation for models using maximum likelihood and Bayes estimators, and (c) Type I error and noninvariance detection rates for scalar measurement invariance models under exact MI, approximate MI, and noninvariance conditions. Based on the noninvariance detection rates, a reasonable cutoff of the prior variance of Bayes estimation was assessed. The impact of simulation factors on the performance of exact and BAMI tests was also evaluated.

Results highlighted that the choice of the prior size affected the BAMI performance, and suggested three pairs of priors for BAMI, (.001 and .05), (.01 and .05), and (.01 and .10), where the first prior in the pair is a representant of approximate-zero invariance while the second prior in the pair is a representant of the substantial non-invariance. In line with the suitable pair of priors, the results also showed that BAMI performed very well if an appropriate fit criterion was used, (e.g., Bayes factor (BF) with 150 as a cutoff and deviance information criterion (DIC)). Implications for BAMI researchers and future directions are discussed.

Chapter One: Introduction

Ensuring measurement invariance (MI) across groups or over time is of particular interest in the psychometrics field because MI is essential in a measure validation (American Educational Research Association, American Psychology Association, & National Council on Measurement in Education, 2014). The measurement invariance concerns about the extent to which the psychometric properties of a scale could be generalized across groups. MI occurred when a participant's observed score depends only on the latent construct score, and has no relation with participant group membership or occasion (Vandenberg & Lance, 2000).

To establish MI, four invariance levels are tested by a set of increasingly constrained models, and differences between these models are evaluated by certain fit indices. The first MI level is configural invariance, which tests whether the scale has the same measurement model across groups. The second MI level is metric or weak invariance, which tests whether participants attribute the same meaning (loadings) to the latent factor across groups. Scalar or strong invariance is the third MI level, which assesses equality of meaning of levels of observed variables across groups. Invariance in residual variances (i.e., strict level) can also be tested as the fourth MI level. It tests whether the unexplained variance of each item is the same across groups (Davidov, Meuleman, Cieciuch, Schmidt, & Billiet, 2014; Kim & Yoon, 2011; Meredith, 1993; Millsap, 2011; Vandenberg & Lance, 2000). If all four MI levels are held, one can conclude that the latent construct is identically measured across groups. However, there is a consensus among researchers that it is not necessary to reach the strict invariance level across

groups (Brown, 2015; Davidov et al., 2015). Demonstrating the advanced MI invariance level, namely scalar, where loadings and intercepts are invariant across groups, is sufficient to validly compare scale scores across groups. MI also allows latent variables to be utilized to hypothesize the relationships among latent variables in structural models (Millsap, 2011; Vandenberg & Lance, 2000).

If a metric or a scalar level does not hold across groups, one can test for partial invariance (Byrne, Shavelson & Muthén, 1989). Partial invariance will be established if there are at least two invariant loadings, (i.e., partial metric invariance), or two invariant loadings and intercepts, (i.e., partial scalar invariance). For detailed information about partial invariance, refer to Byrne et al. (1989).

The majority of educational and psychological assessment research in MI is largely based on multi-group confirmatory factor analysis (MGCFA) using the maximum likelihood (ML) estimation, which is known as the traditional approach. In MGCFA testing for MI across groups, loadings or intercepts are constrained to be equal across groups, where no discrepancies are allowed in measurement parameter estimates across groups. This is not practical when an item parameter is invariant across some groups and non-invariant across other groups if more than two groups are compared (Muthén & Asparouhov, 2013). When the advanced MI doesn't hold, (e.g., scalar), researchers may conduct a sequence of relaxing the non-invariant parameters using ML modification indices, that is, relaxing noninvariant parameters one at a time (i.e., partial invariance). This procedure could be cumbersome and error-prone especially when the number of groups compared is large.

The MGCFA is mainly employed to compare two groups, but it frequently applies to different numbers of groups. MI results using the MGCFA approach across a large number of

groups showed that advanced MI level (i.e., scalar invariance) rarely held. The MGCFA is known for being too strict to meet the advanced MI level, (i.e., scalar). Imposing the exact-zero differences assumption in loadings and intercepts across groups (i.e., identical measurement parameters across all groups) can result in inadequate model fit which leads into inaccurate rejection of the MI level even when the differences are ignorable (Davidov et al., 2015; Muthén & Asparouhov, 2013; van de Schoot et al., 2013). For instance, two common model fit indices are suggested in evaluation of the fit of MI testing (MGCFA) when comparing two groups:1): the change in comparative fit index (CFI) which has a cutoff value of ≤ .01 (Cheung & Rensvold, 2002); and 2) the change in root mean square error of approximation (RMSEA) which should be ≤ .015 (Chen, 2007). These cutoff values, (i.e., ΔCFI and ΔRMSEA) are evaluated to give a judgment of MI testing across two groups. However, when they were used for MGCFA with a larger number of groups, they lead to a frequent rejection of advanced MI because these indices tend to become greater than their cutoff criteria regardless of actual model fit (Rutkowski and Svetina, 2014). Rutkowski and Svetina (2014) stated that a more liberal size of the cutoff value is needed for ΔCFA and ΔRMSEA when using MGCFA to compare 10 or 20 groups, in metric invariance mainly.

An alternative estimation approach was recently incorporated into the context of structural equation modeling. Although it has been more than two decades since ML estimation was the dominant method for MI testing, a Bayesian estimation has become accessible with the availability of different Bayesian software packages. This has caused an increase in the popularity of the Bayesian approach. Researchers have indicated that the ML estimation was dominant due to controversy surrounding the Bayesian approach (Brown, 2015; van de Schoot et al., 2013) and a lack of suitable Bayesian computation software (Brown, 2015; Davidov et al.,

2015; van de Schoot et al., 2013). In addition, Bayesian estimation capability remains applicably ambiguous compared to the traditional approach (Andrews & Baguley, 2012). The view of the anti-Bayesian researchers is that the Bayesian estimation is based on subjective inferences and personal beliefs, which discredits this approach (Andrews & Baguley, 2012). Due to computational advances, a revival of the usage of Bayesian statistics occurred in the late 20th century (Andrews & Baguley, 2012). Most importantly, the increasing demand of Bayesian analysis was due to its capability to solve complex problems (Andrews & Baguley, 2012). Therefore, Bayesian approaches have shown a steady increase in applied research in education, psychology, and social sciences.

Muthén and Asparouhov (2013) suggested a new MI approach, namely Bayesian approximate measurement invariance (BAMI), that utilizes the Bayesian estimation instead of the ML estimation. The BAMI approach relaxes the restrictive assumption of the exact-zero differences in loadings and intercepts variance (i.e., full invariance or exact invariance in MGCFA) and allows for minor discrepancies in measurement parameter estimates across groups by specifying prior distributions of noninvariance (Kim, Cao, Wang, and Nguyen, 2017).

BAMI may solve the issue of over-rejection of scalar invariance. The core of BAMI approach is replacing parameter specifications of exact zeros differences with approximate-zeros based on informative, small-variance priors. In BAMI approach, the minor parameter differences are expected to be zero. Muthén and Asparouhov (2013) argued that this procedure is beneficial in applications. The BAMI approach is efficient in specific settings that traditional MI cannot handle, such as in a large number of groups with many small differences in items loadings and intercepts (Muthén & Asparouhov, 2013; van de Schoot et al., 2013).

Overall, the Bayesian approach can account for uncertainty when modeled properly. Depaoli and van de Schoot (2017) stated that naively applying Bayesian methods may cause certain errors: the influence of priors, misinterpretation of Bayesian features and results, and improper reporting of Bayesian results. BAMI is an innovative method that requires further research to affirm, for example, the prior specifications. Therefore, more research is needed to address the BAMI approach coherently. The present study aims to further examine and explore the performance of the BAMI approach.

Rationale of Study and Background

Measurement invariance (MI) means that participants who have the same ability level (η) are expected to have the same scores for item X regardless of their group memberships ($\mathcal{W}$) (Davidov et al., 2014; Kim & Yoon, 2011; Meredith, 1993; Millsap, 2011; Vandenberg & Lance, 2000). This could be illustrated as in Equation 1:

$$\mathrm{P}\,(\mathrm{X}|\eta) = \mathrm{P}_{\mathcal{W}}\,(\mathrm{X}|\eta). \tag{1}$$

The multiple-group confirmatory factor analysis (MGCFA) is the *most* common way of testing measurement invariance either in cross-national settings or across groups (Millsap, 2011; Vandenberg & Lance, 2000). In MGCFA model, observed item score y_{igj} for individual i, in group g, and item j will be:

$$y_{igj} = \tau_{gj} + \lambda_{gj}\,\eta_{igj} + \epsilon_{igj}. \tag{2}$$

where τ_{gj} is the intercept for item j of group g, λ_{gj} is factor loading for item j of group g , η_{igj} is the latent variable that is to be measured by y_{igj} of individual i in group g, and ϵ_{igj} is the error (see Figure 1).

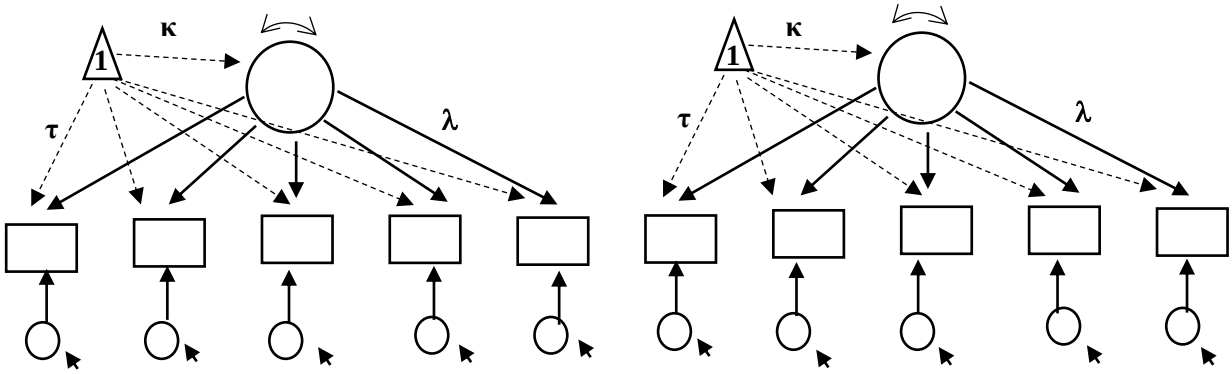


Figure 1. The path diagram of the multi-group confirmatory factor analysis.
Square = observed variable, circle = unobserved variable, curved arrow = variance, short slant arrow = residual variance. κ represents the factor mean; τ represents the intercept of an observed variable; λ represents factor loading. The factor mean of the first group is fixed at zero for identification.

To test for MI, a test of four invariance levels is conducted and in each level, two models are compared: model with invariance constraints and model with such constraints relaxed. The data fit into four sequence models: (1) configural, where groups have the same pattern of confirmatory factor model (CFA) with no equality constraints (i.e., CFA model should be fitted for each group separately which has the same number of factors and same set of zero factor loadings but all other factor loadings and all intercepts are allowed to vary across groups except identification constraints), (2) metric, where factor loadings are constrained to be the same across groups ($\lambda_g = \lambda_{g'}$), (3) scalar, where loadings and intercepts are held constant ($\lambda_g = \lambda_{g'}$, $\tau_g = \tau_{g'}$), and (4) strict, where loadings, intercepts, and residual variances ($\Theta\epsilon_g$) are held constant ($\lambda_g = \lambda_{g'}$, $\tau_g = \tau_{g'}$, $\Theta\epsilon_g = \Theta\epsilon_{g'}$), where $g \; and \; g'$ are two different groups. At each level of MI, the model fit information is compared with the previous one. Scalar invariance is the

required level to be held for mean comparisons and strict invariance is not applicable in many applications (Davidov et al., 2015).

Bayes' Rule: Prior, Likelihood, and Posterior

Bayesian statistical methods allow researchers to apply previous knowledge to new research, which makes this approach unique. Bayesian statistics is not simply another statistical tool; it is a different school of thought, which can tackle model complexity, non-normal data, and small sample sizes (Brown, 2015; Davidov et al., 2015; van de Schoot et al., 2013). In principle, the Bayesian inference is simple, and it has only one tool for coherent inference: the posterior inference (Hoff, 2009; Zyphur & Oswald, 2015). This inference uses Bayes' rule which is the mechanism of the three factors: prior, likelihood, and posterior.

To understand the mechanism of the three ingredients of the Bayes' theorem (prior, likelihood, and posterior), let's say that a parameter, or a set of parameters of interest is θ, y is the observed data, $f(y|\theta)$ is the likelihood (probability density of the data y), and $f(\theta)$ is the prior density. When applying Bayes' theorem to continuous data:

$$f(\theta|y) = \frac{f(\theta)\, f(y|\theta)}{f(y)},$$

$$f(\theta|y) \propto f(y|\theta)\, f(\theta). \quad (3)$$

This indicated that the posterior distribution of θ, given y, is proportional (i.e.,$\propto$) to the product of the probability of y, the data, given θ, the parameter, and the prior distribution (Brown, 2015). Also, Equation 3 could be rewritten to express $f(y|\theta)$ as a likelihood function $L(\theta|y)$ as of Equation 4.

$$f(\theta|y) \propto L(\theta|y)\, f(\theta). \quad (4)$$

Muthén and Asparouhov (2012a) explained Equation 3 in words, as:

$$posterior = parameters \mid data,$$
$$= \frac{data \mid parameters\ X\ parameters}{data},$$
$$= \frac{likelihood\ X\ prior}{data},$$
$$posterior \propto likelihood\ X\ prior. \quad (5)$$

Equation (4) and Equation (5) are essential and represent the core of Bayesian statistics (Brown, 2015; Kaplan & Depaoli, 2012; Muthén & Asparouhov, 2012a; Zyphur & Oswald, 2015).

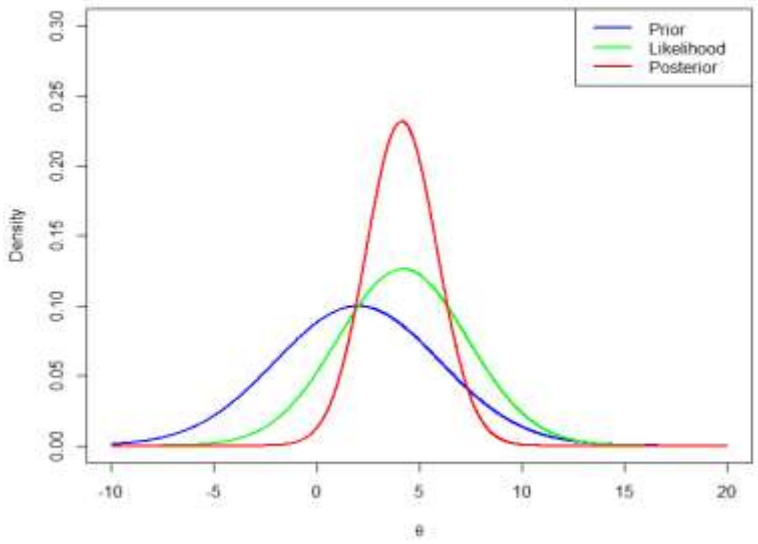


Figure 2. Prior, likelihood, and posterior distributions.
θ= the estimated parameter which could be loading or intercept.

Figure 2 shows the effect of prior, after updated by likelihood, on results of the posterior distribution. It exemplified the concept of the Bayes' rule that displayed a distribution of each of Bayes' three components. The first distribution is the prior of a parameter, θ, which is a likely

value of a parameter based on researcher's knowledge or information in the absence of any data. The second distribution is the likelihood, which is the conditional density of the data given the parameter. The third distribution is posterior, which is a compromise between the prior and the likelihood (i.e., posterior is a product of the prior after updated by the data). A reduction in variability of the parameter estimate (θ) in the posterior distribution compared to the prior, when the data are incorporated, is clearly shown.

BAMI versus Exact MI

In traditional MI testing with MGCFA, fully-invariant parameters or exact-zero parameter differences are assumed. Many studies examining MI across groups (e.g., cross-cultural studies) where there is possibly a large number of groups, criticized the traditional approach using MGCFA for being too cumbersome and showed that the MI assumptions, exact-zero loadings and intercepts differences, are hard to meet at the scalar invariance level in particular (Davidov et al., 2015; Muthén & Asparouhov, 2013; van de Schoot et al., 2013). In other words, the MI assumptions seem strict, and thus need to be alleviated. Also, testing MI with traditional MGCFA is typically suitable for two groups with few non-invariant items (Muthén & Asparouhov, 2013; van de Schoot et al., 2013).

BAMI aims to solve the strict MI requirements issue in the MGCFA, and to make MI more widely accessible. This approach emerged after Muthén and Asparouhov (2012a) suggested the use of the Bayesian approach in structure equation modeling (BSEM). Under structure equation modeling, in the confirmatory factor analysis (CFA) model for a single group, which reflects how the construct is theoretically operationalized, each indicator is allowed to load on one specific factor, while other indicators will have zero loadings with that factor, (i.e., cross-loadings are not allowed; Brown, 2015). The BSEM is a Bayesian approach to analyze

SEM models, especially for cross-loadings and residual correlations in CFA. It allows researchers to test their models with more flexibility by using approximate-zero parameters with zero-mean and small-variance informative priors. Then, Muthén and Asparouhov (2013) and van de Schoot et al. (2013) generalized the BSEM technique by applying the zero-mean, small-variance prior to differences in parameters for testing measurement invariance across groups.

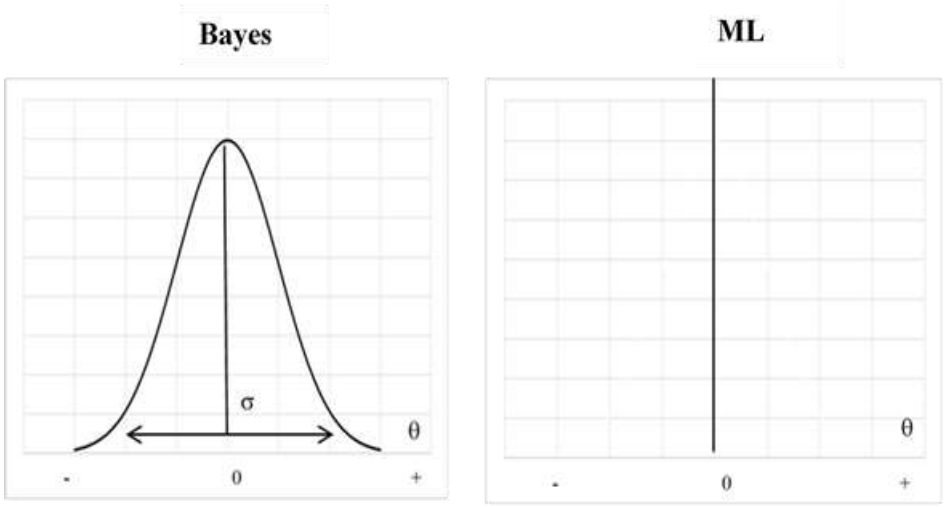


Figure 3. Difference variance in parameter estimation across groups with maximum likelihood and the Bayesian approximate measurement invariance.
ML= maximum likelihood; **σ** = prior variance; **θ**= the estimated parameter loading or intercept.

Many cross-cultural and cross-national studies reported lack of scalar invariance when using traditional MGCAF with ML estimation (i.e., exact-zero differences in loadings and intercepts across groups; Desa, 2014; Nagengast & Marsh, 2013). BAMI was a reaction based on the failure of using traditional MI testing approach. The difference between the ML estimation (exact-zero difference in intercepts or loadings variances) and the Bayes estimation (approximate-zero difference in intercept or loadings variances) is illustrated in Figure 3. The right part of Figure 3 illustrates the ML estimation when difference variances in factor loadings or intercepts, θ, is estimated as exactly zero for CFA models across groups. Using Bayesian

language, ML used a very strong prior with mean zero and zero variance difference in parameter θ. The left part of Figure 3 shows the Bayes estimator using a zero-mean, small-variance σ prior for the difference in parameter θ, (i.e., factor loadings or intercepts variances). It shows the wiggle room (minor discrepancies) that allowed for parameter differences (Muthén & Asparouhov, 2013).

Brief Overview of Bayesian Statistical Inference

In Bayesian inference, three general steps are crucial: the setup of a full probability model, an estimate of a posterior distribution by conditioning observed data, and the evaluation of the assumptions and model fit. Researchers must also be familiar with proper data collection procedures, prior distributions of a parameter θ, the likelihood function, and the conveyed knowledge after data collection (posterior distribution for θ). Bayesian incorporates a prior probability distribution and likelihood of observed data to determine a posterior probability distribution of an event. In other words, a prior distribution is a reflection of the previous information we have about the parameters before confronting the data (Braeken, Mulder, & Wood, 2015). It tells us how to update prior beliefs in light of the new evidence and how to add additional information (Andrews & Baguley, 2012; Hoff, 2009; Zyphur & Oswald, 2015; see Figure 2).

In traditional analysis, ML works by maximizing the data likelihood whereas Bayesian estimation uses prior parameter estimates and then modifies the prior into a posterior (Hoff, 2009; Muthén & Asparouhov, 2012a; Zyphur & Oswald, 2015). The posterior distribution is a compromise between the prior and the likelihood. A key point to differentiate Bayesian from frequentist is the way of viewing the unknown parameter θ. For frequentist, θ is seen as fixed but data are unknown, while in Bayesian, θ (whatever we are uncertain about) has a probability

distribution, and data (whatever you are certain about) are fixed once observed (Andrews & Baguley, 2012; Kaplan & Depaoli, 2012; Hoff, 2009: Zyphur & Oswald, 2015). The most defined distinction in the Bayesian approach is the computation representations, or the summary of the entire distribution (Jackman, 2000).

The BAMI concept implies two steps: to permit replacement of exact-zero variance with approximate-zero variance for differences in parameters (θ) by specifying informative small-variance, and then to relax the non-invariant parameters. In order to permit latent factor means comparison, the BAMI results should show parameters differences across groups that are close to zero (Muthén & Asparouhov, 2012a; van de Schoot et al., 2013). The BAMI approach is best applied when the study has a large number of groups and the scale has many items with small variances in opposite directions (i.e., cancel each other out between groups; Muthén & Asparouhov, 2012a, 2013). When the traditional MI (i.e., full and partial invariance) does not hold given the data, the BAMI could be established for group comparisons (Muthén & Asparouhov, 2012a; van de Schoot et al., 2015). A handful of research has been conducted that diversely addressed the BAMI approaches and applications (see Chapter 2). The BAMI approach is detailed and heavily discussed in Chapter 2.

BAMI Model Fit Criteria

In Bayesian statistics generally and in BAMI specifically, model fit indices and model fit comparison are limited. However, any goodness-of-fit index used for Bayesian statistics could be used for evaluating the BAMI. The posterior predictive p-value (PPP) and the 95% credibility interval (95% CI) for the difference between the observed and the replicated χ^2 values are the two available fit indices to evaluate the fit of a potential model to the observed data (Kaplan & Depaoli, 2012; Muthén & Asparouhov, 2012a). The PPP is the posterior mean that averages over

the posterior distribution whereas the 95% CI gives a range of values on the posterior probability distribution that includes 95% of the true population value. According to Gelman et al. (2014), the cutoff values for PPP are ranged between .01 or .05 and .99. However, the 95% CI χ^2 should encompass zero. There is no indication whether the CI is symmetric or not. Other Bayesian indices are used to compare models such as deviance information criterion (DIC) and the Bayesian information criterion (BIC), where the model with the smaller BIC or DIC value is selected (e.g., Kaplan & Depaoli, 2012; Kim et al., 2017; Muthén & Asparouhov, 2012a). Additionally, Bayes factor (BF) has been used for model comparison in several Bayesian studies, (e.g., Braeken et al., 2015; Kaplan & Depaoli, 2012; Kass & Raftery, 1995; Wagenmakers, 2007). However, no BAMI study made use of the Bayes factor (BF) as a model comparison index in BAMI studies. Detailed information about how to calculate these indices will be discussed in Chapter 2.

Problem Statement

Bayesian statistics accommodate complex models of research since researchers may apply their previous knowledge to new research. This flexibility, along with other reasons, such as dealing with non-normal data and complex models and adhering to small sample sizes, lead to the prevalence of Bayesian statistics in applied research across fields of science (Depaoli & van de Schoot, 2017). However, the Bayesian estimation framework requires a researcher to make decisions throughout the model estimation process. These decisions, sometimes difficult, may affect the estimation process.

BAMI is a promising approach which, when properly utilized, helps a researcher achieve a reasonable and defensible decision regarding measurement invariance from the scientific information already presented. However, Davidov et al. (2015) stated that "the Bayesian test of

approximate invariance cannot establish approximate invariance when measurements are completely different; it does not perform 'magic'. However, it can inform researchers when measurements are sufficiently similar to allow meaningful substantive comparisons" (p. 262). BAMI was implemented in several studies; however, researchers provided limited information in terms of the size of acceptable difference, the use of method procedure, model fit criteria, and the method of interpreting results (Davidov et al., 2015; van de Schoot et al., 2013). Many questions have been raised from applied researchers about the rules of the BAMI process. More importantly, there is no consensus on the best fit indices to evaluate the model fit. There are still some grey areas related to the procedure of BAMI, its use, and the experience needed to use it. Therefore, further research utilizing BAMI is warranted to help researchers make informed choices about applying the best MI approach.

Muthén and Asparouhov (2013) framed the BAMI approach; however, definitive application guidelines have not yet been established. With the current body of literature, several research questions remain unanswered: for example, minimum and maximum thresholds for non-invariant items, bias size and direction, prior variance estimation, and fit indices cutoff criteria. Kim et al. (2017), van de Schoot et al. (2013), and Muthén and Asparouhov (2013) ran BAMI simulations where they provided rough estimates for the prior, cutoff criteria for fit indices, and acceptable level of variance, but more studies need to be conducted to validate these results.

Purpose of Study

Given the lack of knowledge regarding the model fit criteria in the BAMI, it is imperative to examine the behavior of the model fit criteria. To evaluate the feasibility of using the Bayesian approach to conduct measurement invariance testing, the purpose of this study is to examine the behavior of BAMI for use in investigating non-invariance of single level scales that comprised of

continuous items under different design factors. The simulation factors include: number of groups, percent of groups with non-invariant items intercepts, the intercept differences directions, and the magnitude of non-invariance. The research questions are described as follows.

Research Questions

To facilitate the aforementioned study purpose, I conducted a simulation study with similar conditions to those of Kim et al. (2017), Muthén and Asparouhov (2013), and van de Schoot et al. (2013). However, I extended their studies by incorporating additional conditions and analyses of the results through addressing the following questions:

1) What is the performance of the model fit criteria on the BAMI testing in detecting non-invariance level across groups in the single level CFA?
2) What impacts do the design factors (i.e., group number, percent of groups with non-invariant items intercepts, and direction and magnitude of non-invariance) have on the simulation outcomes of testing and estimating the approximate measurement invariance?

Results were examined by observing the behavior of the posterior predictive p-value (PPP), and the 95% credibility interval (95% CI) for the difference between the observed and the replicated χ^2 values. Also, I compared models using the Bayes factor (BF), deviance information criterion (DIC), and the Bayesian information criterion (BIC). Specific criteria for model fit and comparison indices were provided. Supporting the correctly specified model against the competing models was expected by the model comparison indices. The selection rates were summarized across the 100 replications. The simulation outcomes include the proportions of convergence and the detection rates. The detection rates of noninvariance (rejection of scalar invariance) were examined and a description of the impact of each factor on the simulation

outcomes was provided. Moreover, the best fitting models with different level of prior precision values were examined. Discussion in the simulation outcomes can be found in Chapter 3.

Significance of the Study

The BAMI is a newly used approach, which has not been well explored under different research settings (Davidov et al., 2015; van de Schoot et al., 2013). Multiple research teams (Kim et al., 2017; van de Schoot et al., 2013; Muthén & Asparouhov, 2013) conducted BAMI simulations studies where they provided rough estimates for the prior, cutoff criteria for fit indices, and acceptable level of invariance. Yet, more studies are needed to adequately tackle the BAMI and to validate these results. Moreover, Kim et al. (2017) called for more studies to examine the acceptable approximate MI about the magnitude of non-invariance. They stated that:

> It is strongly recommended that applied researchers accumulate knowledge on the magnitude of noninvariance (e.g., reporting the estimates of factor loadings and intercepts and their differences across groups beyond the level of MI; examining the impact of noninvariance in subsequent analyses) with the scales frequently used in their fields to take advantage of specifying priors in the Bayesian approach to MI testing. (p. 16)

We still need research related to how well the BAMI method functions under various research conditions of non-invariance. By varying the magnitude of group differences, one could investigate the extent to which the large group differences must be flagged as non-invariant by the BAMI approach.

Although the typical applications of BAMI are country comparisons, studies also illustrate the BAMI method's potential for handling a different number of groups, such as two groups (van de Schoot et., 2013), 10 groups (Muthén & Asparouhov, 2013), and 25 and 50 groups (Kim et al., 2017). Studies also show how the BAMI method completely automates the task of holding MI across many groups and identifies non-invariant items or groups. This could

change the way large-scale assessment is conducted by replacing the traditional, more arduous techniques, which would make many group comparisons more accessible and valid.

Because the field of approximate MI is relatively new, several questions remain unanswered. Van de Schoot et al. (2013) stated that the two variables that influence the performance of BAMI most are the number of items and magnitude of differences. They suggested future researchers to examine the cutoff values for these decisions (e.g., magnitude of differences acceptable for approximate MI). Because BAMI is a not well established, researchers need to identify and detect violations of the assumptions, and to what extent those violations can affect their results. For example, in the BAMI method, researchers need to know at what number of non-invariant items they may safely perform group means comparison or what acceptable parameters differences across groups can be before results are biased. The main focus of this study, therefore, is to understand under what circumstances that the BAMI method would be optimal. I examined the behaviors of the BAMI method under a variety of magnitudes and directions of non-invariance to inform these types of recommendations.

Limitations and Delimitations

This dissertation has several limitations. Few studies have used the BAMI method, namely 10, yet only three simulation studies (Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al. 2013) were conducted. This shortage of empirical studies, especially the ones not included in the systematic review due to inclusion criteria, may limit the fundamental basis of the study results.

In this simulation study, I purposefully selected certain simulation conditions that reflected various applications of data in education, psychology, and social sciences. However,

some conditions were intentionally avoided for simplicity (e.g., number of items with non-invariant intercept was fixed to 4).

Because models are complicated in simulation studies of measurement invariance, a difference is often made that either loading or intercept non-invariance is simulated (e.g., Cheung & Rensvold, 2002; Kim et al., 2017; French & Finch, 2008; Meade & Bauer, 2007; van de Schoot et al., 2013). For this study, I aim to test for scalar level since it is often the difficult level to be held in traditional MI across groups. It is also the required MI level for conducting the factor mean comparison across groups. Therefore, the metric level was assumed (i.e., invariance of items loadings), and I only tested the models for equality of item intercepts. Accordingly, findings of this simulation study are applicable to the simulation conditions included.

In regard to the study's delimitations, I directly examined the model fit criteria performance with a congeneric single level confirmatory factor analysis (CFA) that employed continuous data. The CFA model included a single factor scale with six continuous items. This scale length was selected according to simulation studies that were done using BAMI (six items in Kim et al., 2017; six items in Muthén and Asparouhov, 2013; four items in van de Schoot et al., 2013). Data were generated under the assumption of multivariate normality. Factor loadings were homogeneous, (e.g., metric level assumed since the scalar level was only of interest). All items loaded on a single factor by varying the factor loadings between .8 and .6 and intercept of zero. The simulation factors were determined based on the results of the systematic review (see Chapter 2).

The overall population was single level CFA model, and I manipulated the model to fit various conditions. Markov Chain Monte Carlo (MCMC) simulation with Gibbs estimation was used. All statistical analyses were performed using SAS 9.4 program and the M*plus* statistical

package (version 8, Muthén & Muthén, 1998-2017). Simulation conditions were associated with the number of groups (GN; medium (8), and large (20)), percent of groups with non-invariant items intercepts (50% and 80%), the intercept differences directions (cancel each other and systematic), and the magnitude of non-invariance (zero, small (.01), moderate (.2), and large (.6)). Two methods were used to test parameter differences (DIF): traditional MI testing with ML estimation to test exact scalar invariance and BAMI to test approximate scalar invariance. The total number of conditions were 2*2*2*3= 24 conditions with DIF items and 2 conditions with no DIF items with 100 replications for each condition.

The PPP and 95% CI were assessed as model fit criteria. For model comparisons, Bayes factor (BF), BIC, and DIC were considered. The PPP and the 95% CI were used in Kim et al. (2017), Muthén and Asparouhov (2013), and van de Schoot et al. (2013). Kim et al. (2017) used the BIC and DIC for model comparison. An investigation of BF as a model fit comparison index was conducted. The conditions of the simulation study were summarized in Table 1.

Table 1
Simulation Study Design

Manipulated Factors	
Number of groups	8, 20
Percent of groups with non-invariant item intercepts	50%, 80%
Magnitude of non-invariance	Zero, small (.01), moderate (.2), large (.6)
Intercept differences direction	Cancel each other, systematic
Constant Factors	
Group size	500
Number of non-invariant items	4 items
Location of non-invariant parameter	Intercept
Factor model	Single factor CFA
Scale length	Six items
Data	Continuous

Definitions of Terms

Maximum likelihood estimation (ML). ML is the traditional procedure to estimate parameters for a given statistic which maximizes the known likelihood distribution.

Bayesian estimation. An estimation method of statistical models where researchers can apply prior information of parameters into their models. Then, parameter estimates are computed based on the posterior distribution of the parameters.

Bayesian structure equation modeling (BSEM). Researchers use the Bayesian estimation to estimate the parameters in the structure equation models via relaxing the exact-zero variance of cross-loadings and using approximate-zero instead.

Prior. Prior is the advanced information that researchers obtain from experts and previous studies in a field about the model parameter.

Posterior probability distribution. The probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey.

Bias. A difference between the generated and the estimated values of a parameter.

Triplot. A single graph to display the prior distribution, the likelihood function, and the posterior distribution. It is used to examine the effect of the data and the choice of prior on any posterior distribution.

Sensitivity analysis. Sensitivity analysis is carried out with multiple plausible prior variances to investigate the robustness of the results of the Bayesian analysis to uncertainty about the precise details of the analysis.

Organization of the Study

This dissertation is organized into five chapters. In the first chapter, I discussed the present approaches for conducting the MI testing across groups, and I identified the need to implement a flexible approach such as BAMI. I also provided the rationale of study and background, Bayes' rule, brief overview of Bayesian statistical inference, the purpose of the study, research questions, study limitations and delimitations, and the definitions of terms. In the second chapter, I synthesized the literature on BAMI testing. I also addressed the gap in the literature and discussed the need to investigate and explore the BAMI approach. In the third chapter, I presented the research method, design, data generation, simulated factors that directed my study, and expected outcomes. In the fourth chapter, results of model convergence, model fit evaluations, detection rates for model fit and comparisons criteria, and assessment of priors were presented. The impact of each of the simulation factor were described and discussed. Finally, in the fifth chapter, a summary of the study and the main findings, discussion, study implications, limitations and direction for future researchers were provided.

Chapter Two: Literature Review

This chapter consists of five parts. First, measurement invariance (MI) across groups is discussed, including definition, applications, and traditional and Bayesian approaches and their challenges. Second, the Bayesian approximate measurement invariance (BAMI) approach is defined and described by addressing the differences between exact and approximate MI, identifying the optimal usage of BAMI, and discussing the advantages, disadvantages, and decisions within BAMI. Additionally, BAMI testing procedures are described, and the corresponding recommendations for each testing procedure are provided. Third, brief operational definitions of Bayesian and M*plus* essential concepts and terminology are provided. Fourth, a systematic review of (10) BAMI applied research is conducted, and the review results are presented and discussed. And finally, three BAMI simulation studies are briefly reviewed regarding their simulation factors and Bayesian decisions.

Measurement Invariance

As stated earlier in Chapter 1, measurement invariance (MI) is required to ensure the validity of using a scale across groups. MI occurs when the measurement model parameters are statistically equivalent across two or more groups/times (Meredith, 1993; Millsap, 2011; Vandenberg & Lance, 2000). Traditionally, MGCFA is used to conduct MI testing across groups. MI is tested incrementally. Depending on the target of the MI levels, (i.e., configural, metric, scalar, or strict), the scale structure, factor loadings, intercepts, and residuals are supposed to be identical across the comparison groups. However, due to diverse issues,

especially in cross-country/cultural research, full or partial scalar invariance is rarely held. If the MI level doesn't hold, MGCFA employs the ML modification indices to relax the non-invariant items, which can lead to a long series of model modifications with substantial risks of misspecification or model rejection of advanced MI level (scalar MI does not hold; Muthén & Asparouhov, 2013). Therefore, testing MI over a large number of groups is methodologically challenging (Kim et al., 2017).

Some researchers (Davidov et al., 2014; Davidov et al., 2015; Muthén & Asparouhov, 2013; van de Schoot et al., 2013) criticized using the multigroup confirmatory factor analysis (MGCFA) for measurement invariance across a large number of groups, mainly for two reasons. First, if the MGCFA is implemented to compare a large number of groups, Type I error (i.e., reject the correct model) could inflate due to a large number of pairwise comparisons across groups. Second, the MGCFA model evaluation goodness-of-fit indices are mainly suggested to compare two groups. Hence, these cutoff criteria may not be appropriate for a large number of groups because the number of groups under comparisons may affect the validity of the criteria results. The criteria might be conservative for a comparison of large number of groups (Kim et al., 2017; Rutkowski & Svetina, 2014).

A new MI approach, Bayesian approximate measurement invariance (BAMI), aims to solve the strict MI requirements issue, the exact-zero variance constraints in loadings differences, in metric MI level, or loadings and intercepts differences, in scalar MI level, and make MI more widely accessible. The BAMI emerged when Muthén and Asparouhov (2013) generalized the Bayesian structural equation modeling (BSEM) specification to be applicable within MI testing, which was a reaction based on the failure of the traditional MI in some circumstances. They stated that by using the BSEM in MI, the exact-zero constraints in parameters differences would

be replaced with approximate-zero based on theory or research. Put differently, in BAMI, constrains in parameter differences and exact-zero, are replaced with approximate-zero differences by specifying prior distributions of non-invariance. The BAMI allows a small none zero variance, approximate-zero, to exist between loadings and/or intercepts differences across groups by specifying a zero mean and small prior variance for these differences. Without using the Bayesian estimation, the exact-zero variance in loadings and intercepts might be difficult to achieve across multiple groups because some items will be invariant across some groups and non-invariant across other groups (parameters discrepancies > 0).

Because the Bayesian approximate MI approach was not well established in terms of the size of acceptable difference, the implementation procedure, model fit criteria, and the interpretation of results (Davidov et al., 2015; van de Schoot et al., 2013), this review aims to explore and evaluate the methodological techniques in conducting Bayesian approximate measurement invariance tests.

Bayesian Approximate Measurement Invariance (BAMI)

The Bayesian approach allows researchers to construct a distribution of plausible values, namely a posterior distribution, using MCMC algorithms to draw random samples from the posterior distribution iteratively. By using the Bayesian approach, parameters are considered random, and uncertainty is combined in the parameter estimator (de Bondt & van Petegem, 2015; Kaplan, 2014; Kim et al., 2017). Additionally, researchers may incorporate their prior knowledge on parameters in data analyses when they specify a prior distribution of a parameter in the method (Kim et al., 2017; Muthén & Asparouhov, 2012a).

In MGCFA, the advanced MI level (i.e., scalar) is achieved when all item parameters, loadings and intercepts, are identical across groups (Kim et al., 2017). The exact constraints of

item parameters (i.e., zero differences) are not usually applicable across many groups which can lead to rejection of the MI, although the differences are minimal. Muthén and Asparouhov (2013) stated that by using BSEM, the exact-zero constraints were replaced with approximate-zero based on theory and research. Muthén and Asparouhov (2013) adapted the BSEM idea to MI testing, which introduced a new type of invariance to the MI testing (i.e., full invariance, partial invariance, approximate invariance). They used BSEM specification in MI tasting as a way to get rid of using exact-zero variance differences in parameters (i.e., loadings or/and intercepts across groups), while obtaining the same information as the MGCFA with ML modification indices. Allowing minor discrepancies between parameters across groups makes the advanced MI testing more attainable. It also helps to reduce the fallacy of rejecting MI (i.e., scalar does not hold), when scale invariance is tested cross-nationally in particular. Approximate zero means that parameters' differences are expected to be zero, on average. The approximate-zero constraint can be applied to the full MI "without relaxing the invariance specification or deleting non-invariant items" (Muthén & Asparouhov, 2013, p.7). Van de Schoot et al. (2013) expressed the parameters differences/discrepancies between groups as "wiggle-room," which is determined based on the prior's degree of precision.

Approximate versus Full Invariance

Applying exact-zero parameters differences, which means the factor loadings or/and intercepts are identical across comparison groups, is the traditional MI approach for full invariance. However, traditional MI testing across many groups, namely MGCFA, is often too strict, and thus, might lead to inaccurate model rejection or a long series of model modifications with substantial risks of misspecification (Asparouhov & Muthén, 2014; de Bondt & van Petegem, 2015; Kim et al., 2017). Although the MGCFA has been utilized for a large number of

groups, many caveats have arisen. For example, Type I error, that is, falsely detecting non-invariance may increase due to a large number of pairwise comparisons across groups (Kim et al., 2017; Rutkowski & Svetina, 2014). Another issue may arise with model evaluation using goodness-of-fit indices. For example, when the identical measurement parameters across all groups are specified, poor model fit may be incorrectly indicated. This might happen because goodness-of-fit indices for MI testing (i.e., $\Delta CFI \leq .01$ combined with $\Delta RMSEA \leq .015$ according to Cheung and Rensvold (2002) and Chen (2007)) are mainly utilized for two groups comparison (Asparouhov & Muthén, 2014; Kim et al., 2017; Rutkowski & Svetina, 2014). These issues (i.e., false rejection of correct invariant model, model misspecification, and poor model fit) can be avoided by employing the BAMI approach. Because the approximate invariance utilizes the Bayesian estimation, the number of groups is less likely an issue. BAMI can handle a large number of groups, and the consequences associated with the number of groups, such as poor model fit, would be resolved.

Approximate versus Partial Invariance

A key difference between partial and approximate invariance is that in the former, only some items parameters are constrained to zero (minimum of two; Byrne et al., 1989) while the rest of the parameters could vary to a great extent because partial equivalence is under the exact-zero framework (Davidov et al., 2015; van de Schoot et al., 2013). However, partial invariance is controversial for many reasons. For example, the source of non-invariance should be located at an item level, and a reference variable should be correctly identified as invariant. Moreover, partial invariance does not accommodate all types of scale structures (e.g., a scale with a single latent factor and three items) because it requires at least two invariant items (van de Schoot et al., 2013; Zercher, Schmidt, Cieciuch, & Davidov, 2015). He and Kubacka (2015) stated that the

partial MI is not suitable when a scale has fewer than five items with a compression of large number of groups (i.e., TALIS scale with more than 24 countries).

Optimal Usage for BAMI

The Bayesian approximate MI approach becomes optimal with a large number of groups and items with small loadings and/or intercepts differences especially when these differences canceled each other within groups, (e.g., -.2 versus .2; Muthén & Asparouhov, 2013; -.01 versus .01; van de Schoot et al., 2013). It is also known to be used with both continuous, (e.g., van de Schoot et al., 2013) and categorical (binary; Muthén & Asparouhov, 2013) data types. The benefits of the BAMI approach becomes prominent when traditional MI tests do not hold given the data, (i.e., MGCFA, Muthén & Asparouhov, 2012a; van de Schoot et al., 2013). Meanwhile, the BAMI approach is less recommended when full invariance holds, or when partial invariance holds with a large size of non-invariance in a small number of parameters (Kim et al., 2017; van de Schoot et al., 2013). Finally, if substantial noninvariance was presented in a small number of parameters across groups, partial MI outperforms BAMI (van de Schoot et al., 2013).

Advantages of BAMI

The approximate-zero approach cannot be achieved by traditional MI with ML estimation because a model with freely estimated factor loadings or intercepts cannot be identified (Brown, 2015; Muthén & Asparouhov, 2013). In BAMI, a prior variance of .01 for factor loadings differences produces 95% trivial non-invariance loadings bounds of negative or positive loadings (i.e., loading values = ± .2). These small differences will not affect the model fit or hinder the comparability between groups regardless of the absence of the absolute invariant (He & Kubacka, 2015; Muthén & Asparouhov, 2013; van de Schoot et al., 2013). Additionally, contrasted with the MGCFA, the number of the compared groups has nothing to do with the

quality of the BAMI results. Finally, if the non-invariant items detected are of focal interest, the BAMI approach will be best serve for this purpose (Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013).

Decisions within BAMI. To apply the BAMI approach, many Bayesian decisions must be made in order to prevent deceptive information (Gelman & Rubin, 1992). Deciding the choice of prior variance and the source of the knowledge of the differences between parameters are the *most* important decisions in the Bayesian approach. Because the precision of the prior determines the wiggle-room, it will reflect on the ability to detect the non-invariance (Muthén and Asparouhov, 2013; van de Schoot et al., 2013). Caution must be taken when calculating a test statistic, which is the ratio of the noninvariance value to the standard error, because both are affected by the increase of prior variance. Muthén and Asparouhov (2013) stated:

> As the prior variance is increased, the non-invariance of a parameter is allowed to be more freely estimated, that is, the estimate can escape from the invariance value to a larger degree. At the same time, the standard error of the parameter increases as the prior variance is increased. (p. 10)

Although Bayesian software (e.g., M*plus* (Muthén & Muthén, 1998-2017), WinBUGS and OpenBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000)) has a default option that can be used, other decisions such as the number of iterations, the spacing between retained iterations of the final analysis, the number of burn-in iterations discarded, and chain and processor values under the MCMC simulation are important (see Bayesian and M*plus* Essential Concepts and Terminology section; Gelman & Rubin, 1992; Raftery & Lewis, 1996). All these decisions affect the quality of the Bayesian results. Bayesian statistics depend on the researcher's level of knowledge and the amount of information provided; therefore, researchers must justify every step used even if they use program default (Depaoli & van de Schoot, 2017).

BAMI Testing Procedures

BAMI testing is a two-step analysis process, where in the first step, researchers identify the non-invariant parameters by using BSEM, and in the second step, free the non-invariant parameters. In order to specify BAMI, one permits replacement parameter specifications (θ) for exact-zeros differences with approximate-zeros based on informative, small-variance that follows a normal probability distribution (Gaussian; $\theta \sim N\ (0, \sigma_j)$). The second step is relaxing the non-invariant parameters. If the BAMI results show cross-group differences in measurement parameters close to zero, then the use of latent factor mean comparison is more meaningful (Muthén & Asparouhov, 2013; van de Schoot et al., 2013). These two steps were implemented differently in the literature resulting in three procedures. The three procedures may be taken in order to apply BAMI: a) test metric and scalar invariance by specifying both in the same step, b) test two levels only: configural and then scalar, and c) the traditional MI testing procedure. The BAMI model identification is the same as the traditional MI model and thus not discussed in this paper. For the first procedure to apply BAMI, researchers test metric and scalar levels by specifying both in the same step. Researchers set informative priors for loadings and intercepts differences across groups. After that, researchers release approximate constraints for loadings or intercepts that are not supported to be approximately invariant by the data (Muthén & Asparouhov, 2013; van de Schoot et al., 2013; Zercher et al., 2015). If the scale has many items with small loadings and/or intercepts differences where these differences canceled out each other across many groups, the first procedure will be suitable to testing for BAMI (Muthén & Asparouhov, 2013).

The second procedure requires researchers to test two levels only: configural and then scalar. With this procedure, researchers rely on previous MI traditional testing results of their

scale that indicate that the scale held the metric or partial scalar invariance across groups. Therefore, for BAMI, they tested for configural invariance using BSEM with informative, small-variance priors for factor loadings and residual covariances. When the configural invariance held across groups, scalar invariance is tested and evaluated by setting different informative priors for the factor loadings and intercepts differences across groups (de Bond & van Petegem, 2015).

The third BAMI procedure is conducted via four steps, which is similar to traditional MI testing using the Bayesian estimator instead of the ML. In the first step, the researcher uses the Bayesian estimation to test for MGCFA without any equality constraint on factor loadings or intercepts, specifically on the configural level. In the second step, the researcher identifies the approximate invariance prior variance for discrepancies in factor loadings and intercepts. Then, in the third step, a series of approximate metric MI models with several prior variances, including the one selected as a cutoff in step two, is performed. A model comparison is conducted, and the best model fit will be selected. In order to achieve the approximate metric invariance, the prior variance in the selected model is supposed to be smaller or equal to the approximate invariance prior variance selected in Step 2. If the BAMI metric holds, Step 4 can be initiated. In this step, researchers repeat the procedure in Step 3 but for the intercept differences. In order to achieve the approximate scalar invariance, the prior variance in the selected model is supposed to be smaller or equal to the approximate invariance prior variance for intercept differences that was determined in advance (Kim et al., 2017).

Recommended Usage of BAMI Procedure

There is no guideline about when to use each approach and why, and researchers are open to use any BAMI approach depending on the purpose of their research and the sufficient information they had about their scales. However, a brief guide is provided with each method.

According to the three procedures for conducting BAMI, specific research scenarios are recommended with each approach. For example, if the scale is newly implemented, and the MI testing has not been studied before, the BAMI sequence MI testing (i.e., the third BAMI procedure) is strongly recommended because it is thorough and allows researchers to test all MI steps. Whereas if the scale was studied before, the second BAMI approach is more suitable to be applied since researchers have sufficient knowledge about the current MI level for the scale. If the researchers have strong informative prior knowledge about the data or the parameters, testing metric and scalar in the same step (i.e., the first BAMI procedure) is more suitable to be implemented. This is also true if the researcher's interest is on testing intercept invariant only., the Again, according to Muthén and Asparouhov (2013) and van de Schoot et al. (2013), the BAMI is preferably implemented when the traditional MI testing (i.e., full or partial invariance) failed to be achieved.

Bayesian and M*plus* Essential Concepts and Terminology

The following terms and concepts are essential in Bayesian analyses and M*plus* and they used throughout this study. Brief definitions are provided to facilitate common understanding for readers.

Markov Chain Monte Carlo (MCMC). MCMC is a simulation summarization technique that has revolutionized Bayesian analysis (Brown, 2015; Link & Eaton, 2012; Raftery & Lewis, 1996). Because the simulation time is of concern, and thus when one uses the MCMC, it is necessary to determine the simulation running time and the set of initial simulation iterations before distribution stabilization, which are also known as *burn-in* numbers (Link & Eaton, 2012). Burn-in numbers are usually the first half of the total iterations that are always discarded when the chain is stabilized because the results in this phase are influenced by starting values (Brown,

2015). Sometimes, a large number of iterations is needed for convergence, but due to the limited space, saving every simulation is not feasible (Muthén, 2010). Because the MCMC chain is dependent, researchers can *thin* the MCMC chain by saving every k^{th} iteration (e.g., third, fifth, tenth, etc.). *Thinning* and *burn-in* are not mandatory practices, but both help to reduce the amount of data saved when running MCMC (Link & Eaton, 2012; Raftery & Lewis, 1996). Further literature related to MCMC sampling includes Gelman, Carlin, Stern, and Rubin (2014), Kaplan (2014), Hoff (2009), and Spiegelhalter, Best, Carlin, and van der Linde, (2002).

Gibbs sampler. Gibbs sampling or a Gibbs sampler is the MCMC algorithm for obtaining a sequence of observations, which are approximated from a specified multivariate probability distribution when direct sampling is difficult (Brown, 2015). The Gibbs sampler begins with an initial set of starting values for the parameters θ, and given this starting point, the Gibbs sampler generates new θ from the previous one. Then, a sequence of dependent vectors is formed (Brown, 2015). Under some general conditions, the sampling distribution resulting from this sequence will converge with the target distribution (see Hoff, 2009).

Model convergence. Convergence is the key in Bayesian analysis. The convergence of posterior parameters means that the parameters estimate is accurately achieved through a sufficient number of drawn samples (Kaplan & Depaoli, 2012). There is not a specific evaluation criterion of convergence, and hence, researchers use several diagnostics methods and fits. One of them is the *Potential Scale Reduction* (PSR). The PSR is assessed via monitoring of the posterior distributions (de Bondt & van Petegem, 2015; Gelman & Rubin, 1992; Gelman et al., 2014; Kaplan & Depaoli, 2012). Because MCMC could be multiple chains, PSR is used to compare the parameter estimates for within- and between-chain variations. If a single MCMC chain is used, the within and between variations of the third and fourth quarters of the iterations are compared

via PSR. The PSR value 1 represents perfect convergence. However, if a model has a large number of parameters, a PSR value of less than 1.1 for each parameter represents the convergence (de Bondt & van Petegem, 2015; Kaplan & Depaoli, 2012; Muthén & Muthén, 1998-2017).

In M*plus*7, it stops the Bayesian algorithm when the PSR drops below 1 plus very small value between .05 to 1, according to Asparouhov and Muthén, (2010). However, the option "Bconvergence" could be used with a strict cutoff, (e.g., .01), for convergence (Muthén & Muthén, 1998-2017). The assessment of the Bayesian model convergence is difficult due to the design of the MCMC algorithm because MCMC converges in distribution shape rather than a point estimate (Kaplan & Depaoli, 2012).

Biterations. This option is used in M*plus*7 under the "analysis" command in order to specify the maximum or minimum number of iterations for each chain of the MCMC procedure with combination with the Gelman-Rubin PSR convergence criterion. For example, when BITER=50000 (20000) is specified, the MCMC runs for a minimum of 20000 iterations and a maximum of 50000. If the number of iterations reached (20000), the convergence is again assessed using the Gelmin-Rubin PSR criterion (Muthén & Muthén, 1998-2017).

Fbiter. This option is used in M*plus*7 under the "analysis" command. Fibter is used to enable the researcher to manually specify a fixed number of iterations for each MCMC chain when Gelman-Rubin PSR is not used (Muthén & Muthén, 1998-2017).

Bseed. Because MCMC procedures are based on random sampling from the prior and posterior distribution, one may get slightly different results every time the analysis is run especially with different computers. The "Bseed" option is used to specify a random number generation in the MCMC algorithm. Hence, if a value for the Bseed is given, the same random

values sequences will be obtained and results will be always the same. This option also would be useful when models do not reach convergence, so changing the Bseed value to start the MCMC process might help. M*plus*7 default is zero for the "Bseed" option.

Trace and autocorrelation plots. A trace plot shows the history of a parameter value across iterations of the chain. The autocorrelation is between the samples returned by the MCMC. Autocorrelation ranges between -1 and 1, and measures how linearly dependent the current value of the chain is to past values. A check for the trace and autocorrelation plots of the posterior distributions could be used for model diagnostic to judge the convergence (Muthén, 2010). Sampled parameter values over time are presented via trace plots where quick up-and-down variations and absence of long-term trends show quick distribution convergence. Autocorrelation should become smaller as the sampling number increases (de Bondt & van Petegem, 2015; Kaplan & Depaoli, 2012). It should not show any long-term trends. In the MCMC chains, convergence occurs when the degree of correlation for parameter values across iterations (non- independence) measure close to zero (0.1 or lower; Kaplan & Depaoli, 2012; Muthén, 2010).

Model fit evaluation and model comparison. As a means of evaluating the quality of the Bayesian model fit, two main fit indices are popular, particularly in M*plus7*: posterior predictive check with posterior predictive p-value (PPP) and the 95% credibility interval (95% CI) for the difference between the observed and the replicated χ^2 values. Other fit indices for model comparison are the Bayesian information criterion (BIC). Bayes factor (BF), and the deviance information criterion (DIC).

Posterior predictive check and posterior predictive p-value. The posterior predictive check (PPC) accounts for uncertainties in model parameters in data (Gelman, Meng & Stern,

1996; Kaplan & Depaoli, 2012; Muthén & Asparouhov 2012a). It is an index to measure the extent of accuracy generated by the model or the replicated data matched the actual data. This quality of the predictive accuracy by measuring the discrepancy or the deviation is the essence of the *PPC* because it is an indication of possible model misspecification (Kaplan & Depaoli, 2012; Muthén & Asparouhov, 2012a).

Posterior predictive p-value (PPP) is a guide for sensitivity of the prior choice. Muthén and Asparouhov (2013) stated, "If the prior variance is small relative to the magnitude of non-invariance, PPP will be lower than if the prior variance corresponds better to the magnitude of non-invariance" (p. 21). When the model is misspecified, an extreme PPP value is expected (e.g., PPP < .05 or .01; Kim et al., 2017). An extreme PPP value means that the PPP does not belong to the distribution of the correctly specified model and it is in the tail of the distribution. In equation language and from Bayesian theorem:

$$f(\theta|y) \propto f\ (y|\theta)\ f\ (\theta). \tag{6}$$

where a parameter or a set of paramours of interest is θ, y is the observed data, $f(y|\theta)$ is the likelihood, $f(\theta)$ is the prior density. If y^* is the data replicated and as Equation 6, $f\ (\theta|y)$ is the posterior distribution of the model parameter, the probability of future observation given the current data $(y^*|y)$ is the same as probability of future observation given parameters (Hoff, 2009; Kaplan & Depaoli, 2012). Therefore:

$$f\ (y^*|y) = \int f\ (y^*|\theta)\ f\ (y|\theta) d\theta,$$
$$f\ (y^*|y) = \int f\ (y^*|\theta)\ f(\mathrm{y}|\theta)\ f\ (\theta)\ d\theta. \tag{7}$$

PPP is not as usual as the χ^2 test of model fit in traditional statistics, and the value around .5 would be favorable as an indication of an excellent fitting model (Muthén & Asparouhov,

2012a). In addition, according to simulation studies, PPP values of 0.01 and 0.05 are considered sound. (Muthén & Asparouhov 2012; van de Schoot et al., 2013). In summary, the more prior specification is added to the model, the smaller the PPP becomes. Thus, compared to frequentist method, researcher becomes more certain about the results after confronting the prior knowledge with the data (van de Schoot et al., 2014).

Credibility interval (also known as a posterior probability interval). The Bayesian credibility interval (CI) has a different interpretation than the frequentist confidence interval. The latter is based on the assumption of a very large number of repeated samples from the population, while the former, the Bayesian is based on sampling from the posterior distribution. Therefore, it is easy to use the distribution quantiles, and imply that the probability of a parameter lies in the interval (0.95; Kaplan & Depaoli, 2012). The CI is calculated using Equation 8 where a 100 $(1-\alpha)$ % CI for the parameter space θ is:

$$1-\alpha = \int f(\theta|x)\, d\theta. \tag{8}$$

Muthén and Asparouhov (2012a) indicated that the term *significant* in Bayesian is used when the 95% of the CI does not include zero.

The Bayesian information criterion (BIC; also called the Schwarz criterion). Another popular measure comparing models is BIC, which is used with un-nested models. Several statistical packages calculate BIC, such as M*plus7*, OpenBUGS, and SAS based on Equation 9:

$$\text{BIC} = -2\log(\theta|y) + \text{q}\log(n). \tag{9}$$

In this equation, $-2\log(\theta|y)$ is the model fit, q is the parameters number, *n* is the sample size. After calculating BIC for each model, a smaller number indicates better results (Kaplan & Depaoli, 2012).

Bayes factor (BF). Under the posterior predictive checking, a framework of model choice is a key idea in Bayesian statistical modeling considering that the model will be used for prediction (Kaplan & Depaoli, 2012). BF, a Bayesian model comparison tool, is the standard Bayesian measure of relative evidence between two competing statistical models. In essence, BF is used to quantify the odds that the data favoring one hypothesis/model over the other (Braeken et al., 2015; Kaplan & Depaoli, 2012; Kass & Raftery, 1995). The BF is often interpreted as the weight of evidence coming from the data since a BF hypothesis test selects the hypothesis under which the observed data are most likely (Wagenmakers, 2007). Recently, there has been an increasing interest in the use of the BF. Kass and Raftery (1995) defined the BF as the ratio of the marginal likelihoods under two hypotheses or models of interest, where the marginal likelihood provides a Bayesian measure of the support in the data for each hypothesis. Kaplan and Depaoli (2012) defined BF as "the ratio of the posterior odds to the prior odds" (p. 655).

BF has a straightforward interpretation as the relative support in the data between two hypotheses. It allows a comparison between any two models, even models that are complex, nested, and non-nested (Braeken et al., 2015; Kass & Raftery, 1995; Kaplan & Depaoli, 2012). Researchers, especially psychologists, favor using the BF over other model comparisons indices, such as p-value in traditional testing, since BF has an intuitive interpretation where the researcher can claim evidence in its favor and gives more information to inform the decision. For example, MI testing was used as an application of using BF by Verhagen, Levy, Millsap, and Fox (2016). They stated that interpretation of the BF tests for MI was straightforward. See Mulder and Wagenmakers (2016) and Wagenmakers (2007) for more information regarding differences between BF test and classical significance tests in traditional testing especially in psychological research.

BF relies on the full Bayesian approach where each model is given a prior probability which, when multiplied by the marginal likelihood, yields a quantity that is proportional to the posterior probability of the model (Gelman et al., 2014). Although many researchers stress the need to use the BF within a fully Bayesian approach, Gelman et al., (2014) limited the use of BF to specific situations. Gelman et al. (2014) stated that "the marginal likelihood is highly sensitive to aspects of the model that are typically assigned arbitrarily and are untestable from data" (p.182). However, Kaplan and Depaoli, (2012), Kass and Reftery (1995), Muthén and Asparouhov (2012b), and others advocated for the use of BF as an index for comparing models with continuous variables.

Different methods were found in literature in order to compute the BF. For example, BF can be computed as the ratio of prior predictive probabilities times the prior odds (Wagenmakers, 2007). Although most experimental psychologists make the use of the Bayesian Information Criterion (BIC) only to compare non-nested models, it is an easy and quick way to compute BF (Muthén & Asparouhov, 2012b; Wagenmakers, 2007). The idea behind that is the comparison models are equally plausible a priori; therefore, comparing their BIC values easily yields an approximation of their posterior probabilities. Several statistical packages provide BIC values such as M*plus* 7 and up. BIC is included for all models with continuous items in single level models in M*plus* 7. Wagenmakers (2007) provided Equation 10 as a way to compute BF using BIC where the two competing models were labeled as H_0 and H_1 (Wagenmakers, 2007):

$$\text{BF} = \frac{P\,(H1)}{P\,(H0)} = exp\,(\frac{\Delta\, BIC_{H0H1}}{2}) \qquad (10)$$

where $\Delta\, BIC_{H0H1} = BIC_{H0} - BIC_{H1}$. As mentioned by Wagenmakers, (2007) the BF calculation in Equation 10 involves the difference between the two BIC values not the absolute BIC values.

There is no clear adherence to adequate rules for the interpretation of the size of the BF; however, using Equation 6, Kaplan and Depaoli (2012), Kass and Raftery (1995), Muthén and Asparouhov (2012b) and Verhagen et al. (2016) stated that a BF value greater than 3 is considered evidence of supporting H_1. It means that the data is three times more likely under H_1 than under H_0. More classification scheme of the BF can be found in Kass and Raftery (1995). Finally, some researchers use the log of the BF rather than the BF since the log is more stable than the BF when the BF value is very small or very big (Christensen, Johnson, Branscum & Hanson, 2011). In the case of using the log BF, the positive values favor H_1 while the negative values favor H_0.

The deviance information criterion (DIC). Whereas BIC is used in traditional and Bayesian applications, DIC is based on Bayesian deviance and is particularly useful in Bayesian model selection (Kaplan & Depaoli, 2012; Spiegelhalter et al., 2002). The number of parameters used to penalize for model complexity with the DIC is the effective number of parameters, referred as pD. Models with smaller values of DIC should be preferred. DIC estimates the effective number of parameters, and the smaller is the better. Spiegelhalter et al. (2002) stated that DIC is equal to goodness of fit plus complexity. The goodness of fit is measured by the deviance D (θ), and defined as in Equation 11:

$$\mathrm{D}\,(\theta) = -2\log[f\,(y|\theta)]. \tag{11}$$

In this equation, θ is model parameter,$f\,(y|\theta)$ is the likelihood (Kaplan & Depaoli, 2012). While complexity measured by the P_D, which is the estimate of the effective number of parameters as in Equation 12:

$$P_D = E_{\theta|y}\,[D] - D\left(E_{\theta|y}\,[\theta]\right),$$

$$= \overline{D} - D\ (\bar{\theta}). \tag{12}$$

where $E_{\theta|y}$ is the expectation of θ (Spiegelhalter et al., 2002). In other words, the effective number of parameters is the posterior mean deviance ($\overline{D}$) minus deviance evaluated at the posterior mean ($D\ (\bar{\theta})$). Therefore, DIC is defined as in Equation 13:

$$DIC = D\ (\bar{\theta}) + 2\ p_D,$$

$$= \overline{D} + \ p_D\ . \tag{13}$$

Systematic Review for BAMI Applied Research

This review is driven by the need to delineate guidelines of the approach of exploring BAMI method (e.g., set up a prior), number of items that are non-invariant, and how large the typical difference should be appropriate (i.e., the acceptable bias size). Most importantly, researchers need to define their "approximate-zero", "small", or "minor" parameter difference and set up the "golden" rules for evaluating model fits. In this review, researchers' transparency in reporting of their research process and findings will be discussed.

Frameworks

A systematic review of Bayesian articles in psychology conducted by van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, and Depaoli (2017) and BSEM Approximate Measurement Invariance (Muthén & Asparouhov, 2013), along with the Standards for Educational and Psychological Testing by American Educational Research Association [AERA], American Psychology Association [APA], & National Council on Measurement in Education [NCME], (2014) were used as frameworks to develop this review.

The purpose of this systematic review is to evaluate studies that utilized the BAMI through addressing the following research questions:

1) How approximate measurement invariance was tested using Bayesian structure equation modeling? This question will be answered in terms of: a) the procedures or steps followed, b) model identification, c) prior setup, d) model estimation, e) define the "approximate", and f) sensitivity analysis.

2) How BAMI model was evaluated? This question will be answered in terms of: h) convergence and i) model fit evaluation.

3) What level of invariance was achieved? This question will be answered in terms of: j) reported level of invariance before and after BAMI, k) number of measurement non-invariant items.

Search Strategy

An electronic systematic search was conducted to identify studies published between 2013 and 2017. The year 2012 is chosen because the BAMI approach was introduced first by Muthén and Asparouhov in (2012a). Then, to determine the included articles in this review, a set of inclusion criteria would be applied to the found articles.

Search terms and databases. The following databases were used to identify articles: PsycINFO, Education (full text), EBSCOhost, and Google Scholar. The keywords used in the searching were "approximate measurement invariance", "Bayesian measurement invariance", "Bayesian approximate invariance", "Bayesian multiple-group measurement invariance", and "approximate Bayesian measurement invariance across groups". The initial search yielded 90 articles. From the initial pool, 40 articles were excluded because they were replicated on several datasets, and 50 articles remained for second search round. In the second round of search and from the 50 articles, 24 articles were excluded because the keywords merely appeared in the texts or references of the articles but BAMI was not actually used in the study. Reference lists for

the remaining 26 articles were checked for additional applications. The articles were sorted into three categories: applied studies were built on the traditional MI results for the same measure (40%), applied studies for the purpose of validation of a new measure (40%), and studies used BAMI as a demo part to apply simulation results (20%).

Inclusion Criteria

After the database search, I applied the inclusion criteria to the remaining articles. To include a study in this review, the study must: (1) have model of multigroup single confirmatory factor analysis (CFA) or item response theory (IRT), (2) use Bayesian approximate measurement invariance approach to address major research question, (3) be published in between 2013 and 2017, (4) and in case of simulation study, the demo section will only be used. Table 2 represents a summary of the inclusion and exclusion criteria.

Table 2

Review Inclusion and Exclusion Criteria

Review Criteria	Inclusion	Exclusion
Date	From 2013 to 2017	After 2017
Type	Scholarly articles	Book chapters, reports, dissertations, proceedings.
Design	Applied	Simulation
Research Question	BAMI as main research question.	Other
Model	Multigroup (CFA or IRT)	Multiple level CFA or IRT
Method and results	Identified method, described results section	Reviews, conceptual paper, reports, or discussion.
Language	English Journal article	Article written by Other languages
Estimator	Bayesian	Traditional

Note. CFA: confirmatory factor analysis; IRT: Item response theory.

Assessment of BAM Usage

Review Protocol (Method of Analysis)

The coding instrument was developed to identify and record the key parameters in order to answer the four research questions. It was developed in stages and was based on several

sources including a review of the (a) literature on technical and methodological issues related to Bayesian, (b) journal articles discussing statistical reporting practices, including the AERA et al. (2014) Standards for Educational and Psychological Testing, and (c) discussions between Bayesian statisticians and myself.

The coding protocol had 57 specific items addressing three major areas: a) approximate measurement invariance testing procedure in terms of model specification and estimation (i.e., MGCFA model, model identification, MI level before BAMI, number of bias items before BAMI, number of model tested, BAMI procedure, algorithm, prior, prior distribution for factor means and variances, prior distribution difference in loadings/ intercepts/ residuals variances, prior for residual covariances, residuals correlated/ uncorrelated); b) approximate measurement invariance model evaluation in terms of convergence and model evaluation (i.e., convergence inspection, fit indices, model evaluation); and c) level of invariance achieved in terms of the results (i.e., level of MI hold, number of bias items, results presentation). The option "N/A" is available when that element is not reported or not discussed. See Appendix A in the appendices for a copy of the BAMI coding protocol for this review.

Inter-Rater Reliability

To measure the extent to which a data collector records the same scores for the same item, the Cohen's kappa inter-rater reliability was used because the percent agreement is unable to account for chance agreement (Cohen, 1960). Cohen's kappa ranges from -1 to +1 (McHugh, 2012). Because this review was conducted by a single coder, to avoid subjectivity, two of the articles were independently coded by a psychometric professor and myself. Inter-rater Kappa for multiple raters was computed using Stata version 13.1 (2013). Kappas for two articles were .93%, and .97%, and the overall kappa was .95%. For disagreements issues, we discussed the

disagreements until we reached to consensus. After establishing the inter-rater reliability from the previous step, I followed two-step inter-rater reliability method (Mackey & Gass, 2005). Firstly, I coded all the data, and after some lapse of time (two weeks), I recoded the data, which means the coding was done by single coder but at different times. Next, I compared the scores using Kappa. The Kappas for the 10 articles ranged from 92% to 96%.

Systematic Review Results

After reading the full text, I excluded 16 articles out of the 26 articles that used MI Bayesian estimation because these articles used different Bayesian MI techniques such as alignment or random effect modeling. Appendix B showed the articles alphabetically ordered by title with their assigned number, authors, year, and journal. Also, see Appendix C for the *PRISMA Flow Chart* for the citation process. For a list of the full citation of reviewed articles, see References section for reference with an asterisk. Results are organized in the same order of the systematic review research questions. Each section answers one of the systematic review research questions.

Approximate Measurement Invariance Testing Procedure

BAMI procedure. As explained earlier on the BAMI procedures, the three BAMI procedures were observed. Six studies (60%) tested metric and scalar at the same step, so they relied on previous research results or current traditional MI levels. The second approach was adopted by only one study, where it tested two levels only: configural and then scalar. The third approach, which was adopted by 30% of the studies, used the sequence MI testing procedure where they tested for configural, then metric, and then scalar, or proceeded into approximate scalar if needed (see Table 3).

Model identification. As stated before that the BAMI model identification is the same as the traditional MI model either by using marker-variable or by standardized the factor. Both approaches were used within the 10 studies. Forty percent used the marker variable (Cieciuch et al., 2014; Davidov et al., 2015) and 60% of studies fixed the factor variance at 1.0 (de Bondt & van Petegem, 2015; Muthén, & Asparouhov, 2013; van de Schoot et al., 2013).

Table 3

BAMI Procedures across the Reviewed Articles

Study Assigned Number	One Step: Scalar and Metric	Two-Step: Configural then Scalar	Three-Step MI
1		X	
2	X		
3			X
4	X		
5			X
6	X		
7			X
8	X		
9	X		
10	X		
Total	6	1	3

Note. X= Study used the corresponding procedure; BAMI= Bayesian approximate measurement invariance; MI= measurement invariance.

Prior specification. The 10 studies included in this review followed their frameworks in quantifying priors. Factor mean and variance priors were specified as noninformative (diffuse) or normal prior distribution for loadings and intercepts with prior mean of zero and variance of 10^{10} by three studies (30%; Gucciardi et al., 2016; van de Schoot et al., 2013; Zercher et al., 2015). Others stated that the prior distribution for loadings and intercepts are freely estimates (e.g., Braeken & Blömeke, 2016; Davidov et al., 2014). Several articles did not specify the model clearly but provided the option "model = allfree" in M*plus* syntax (e.g., Bujacz et al., 2014; Cieciuch et al., 2014). This M*plus* option means that all model parameters (i.e., loadings, intercept, and residuals) are free (Muthén & Muthén, 1998-2017) except for those for identification purposes (i.e., marker variable). Twenty percent of the studies did not mention the

factor mean and variance prior distribution (He & Kubacka, 2016; Muthén & Asparouhov, 2013). Thirty percent specified prior residual covariance as: inverse gamma distribution (-1, 0; van de Schoot et al., 2013), (0, .006; Gucciardi et al., 2016), or just mentioned noninformative covariance (de Bondt & van Petegem, 2015).

On the other hand, all studies (100%) specified prior distributions for the differences in loadings and intercepts. Two types of prior specifications were found: (a) priors of intercepts and priors of factor loadings are the same and (b) priors of intercepts and priors of factor loadings are different. Across 90% of the studies, prior distribution differences in loadings and intercepts for all items were assigned the same. Only one study (10%) specified two difference priors for its factors item intercepts. Namely, Cieciuch et al. (2014) study, which has a scale of 19 factors, and 16 of them have informative prior with normal distribution loadings and intercept difference with mean of zero and variance of .01 (i.e., N~ (0, .01)), whereas the last three factors have informative normal distribution for different loadings and intercept with mean of zero and variance of .02 (i.e., N~ (0, .02)). Those priors were normally distributed and informative with small variances. Different researchers also addressed prior knowledge and they also justified their prior options. For example, Bujacz et al. (2014) stated that they picked the N~ (0, .01), normal distribution with mean of zero and variance of .01 for difference on loadings because it will allow approximate-zero for factor loadings, but will keep them small and insignificant.

Different priors' loadings and intercepts variances were specified for the same model, (e.g., 0.005, 0.01, 0.05, 0.10, 0.2, 0.5). Although 60% of the studies used four or five different priors for the same model, they used priors that were proposed by Muthén and Asparouhov (2013) and van de Schoot et al. (2013). Seventy percent of studies used either N~ (0, .05) or N~ (1, .01) as the best option with which they determined the measurement invariance level. Thirty

percent of studies used only one prior, namely .05 or .01, and only 10% used eight priors from .10 to a very extremely small variance .00000001(i.e., de Bondt & van Petegem, 2016). Priors reported across the studies were: N~ (0, .0005), N~ (0, .005), N~ (0, .05), N~ (0, .02), N~ (0, .01), N~ (0, .10), N~ (0, .2), and N~ (0, .5) by 20%, 20%, 70%, 10%,70%, 20%, 10%, and 20% of the studies, respectively. Again, the N~ (0, .05) and N~ (0, .01) were the most reported priors values across the results. Table 4 provides a summary of each study with their corresponding prior.

Table 4

Summary of Reported Prior per Study

Study	Prior Variance for Loading and/or Intercept								Total
	.0005	.005	.05	.02	.01	.10	.2	.5	
1[a]					X				8
2			X						1
3					X				1
4				X	X				2
5			X			X			2
6	X		X		X			X	4
7		X	X		X				3
8	X	X	X		X			X	5
9			X						1
10			X		X	X	X		4

Note. [a] Study (1) used eight priors ranged from .1 to .00000001.

Another source of reported priors was BAMI simulation studies, (namely, Muthén and Asparouhov, 2013; van de Schoot et al., 2013). Those studies were the source of prior knowledge for 50% of the reviewed articles (e.g., Cieciuch et al., 2014; Davidov et al., 2015; de Bondt & van Petegem, 2015; Zercher et al., 2015). Because van de Schoot et al. (2013) recommended the use of several prior variances and then compared the results, the same scenario occurred with several studies later on (e.g., Gucciardi et al., 2016; He & Kubacka, 2016; Zercher et al., 2015). Only one study (10%) specified the source of prior as the MGCFA traditional result (Braeken & Blömeke, 2016). Eighty percent of the studies have not reported correlation between

factors or errors. Because one study has correlated errors (Braeken & Blömeke, 2016) and the other one has correlated factors (Cieciuch et al., 2014), they counted them as a part of their models because the models provided poor fit without these covariances. Appendix D showed Wordcloud presenting terms used to describe the level of informativeness of the priors in the review.

Model Estimation

Because 100% of the articles used Mplus program to conduct the Bayesian estimation, they used MCMC algorithm. Forty percent of the studies used Gibbs sampler, the Mplus default, to estimate the Bayesian model (e.g., de Bondt & van Petegem, 2015; Gucciardi & Zhang, 2016), whereas the other 60% did not report. However, 60% of the studies has supplementary materials (e.g., Bujacz et al., 2014; de Bondt & van Petegem, 2015; Davidov et al., 2015), and 50% of them included the Mplus code. Twenty percent of the studies provided the Mplus code as a part of the study (e.g., Cieciuch et al., 2014), whereas only three studies (30%) (e.g., Gucciardi et al., 2016) described some of their Bayesian codes (e.g., chain and iteration numbers within the study).

Across studies, several M*plus* code options were shared. Forty percent of studies (Bujacz et al., 2014; Cieciuch et al., 2014; Zercher et al., 2015), for example, had "biterations" values, which ranged between a maximum of 200000 and a minimum of 20000. Three studies (30%) (e.g., Gucciardi et al., 2016) used fixed iteration numbers of 150000 for each chain, several studies (e.g., van de Schoot et al., 2013) used 5000, and the other used 1000 (e.g., de Bondt & van Petegem, 2015) with "Fbiter" M*plus* option.

MCMC "chain" number by M*plus* default is two (Muthén & Muthén, 1998-2017), but chain values differ across studies. In all cases, however, processor and chain numbers were

matched. For example, for both numbers of chains and processors, four studies (e.g., He & Kubacka, 2015; Muthén & Asparouhov, 2013; 40%) used four chain values, one study (i.e., Zercher et al., 2015) used eight, another study (i.e., de Bondt & van Petegem, 2015; 10%) used two, Cieciuch et al. (2014) used five, and only 30% of the studies did not report the chain values. Finally, studies had different Bseed values such as 100 (Cieciuch et al., 2014; Zercher et al., 2015), 20 (van de Schoot et al., 2013), and 200 (He & Kubacka, 2015).

Define approximate. For the "approximate-zero" or "approximate equality", it is important to describe how large the permissible differences on loadings or intercepts are. The quantification definition of "approximate-zero" was ignored in most of the studies, and different terms were used such as "vary slightly" by Davidov et al. (2015), "small differences" by He and Kubacka, (2015) and Zercher et al. (2015). However, the information on selected prior variances was provided (see Table 4). Bujacz et al. (2014) stated that they picked the .01 prior because the difference will be small and insignificant and ranges between -.2 and +.2. Muthén and Asparouhov (2013) estimated the prior of .10 because they believed that 95% of the distribution of the non-invariance lies between ± .62. They also stated that 95% of the distribution of the non-invariance difference lies between -.22 and .22, when prior variance is equal to .01. These were only the definitions provided in all studies.

Sensitivity analysis. As aforementioned, results of Bayesian analysis are sensitive to any change. Therefore, conducting a sensitivity test, with multiple plausible prior variances, is recommended. For example, a sensitivity analysis was conducted to investigate "the effects of varying the prior variance of the residual covariances on the PPP and the lower and upper bounds of the 95% CI for the difference in chi-square statistic for the observed and synthetic data" (de Bondt & van Petegem, 2015, p.9). It is also used to check the variability of the estimated

parameters, the results of the BAMI are supposed to be approximately the same and don't alter the estimation of the parameters considerably, unless the sample size is extremely small and/or the model or prior distribution is strongly contradicted by the data (de Bondt & van Petegem, 2015). Although most of the studies (90%) recommended carrying out sensitivity analysis, only three studies (30%) (e.g., van de Schoot et al., 2013) did, and two of these three studies had also simulation data (e.g., Muthén & Asparouhov, 2013).

Approximate Measurement Invariance Model Evaluation

Convergence

Convergence was assessed visually and statistically. Forty percent of the studies discussed the convergence cutoff and the use of the potential scale reduction *(*PSR) as a criterion, which is approximated to the value of 1 as a cutoff (e.g., Muthén & Asparouhov, 2013). However, terms used in studies were vague such as approximate one (Braeken & Blömeke, 2016) and around one (Muthén & Asparouhov, 2013). Also, other PSR values such as 1.01 and 1.05 were used by only one study (de Bondt, & van Petegem, 2015) that has a big sample size. Eight studies did use the visual inspection for the MCMC trace and autocorrelation plots (e.g., de Bondt & van Petegem, 2015; Gucciardi et al., 2016). However, only 20% of the studies (e.g., de Bondt & van Petegem, 2015) provided the actual trace plots.

Model Fit and Model Comparison Indices

The 95% credibility interval of χ^2 (95% CI) as well as PPP were the main model fit indices across all studies. Fifty percent of the studies used the value of PPP <.05 as an indication of a poor fit model, and 30% of them indicated that PPP should be above zero, while 20% of the studies indicated that PPP should be greater than .01. In addition to model fit indices, (i.e., PPP and 95% CI), model fit comparison indices were provided. Twenty percent of the studies used

the deviance, 30% used deviance information criterion (DIC). However, no study utilized the Bayes factor (BF) or the Bayesian information criterion (BIC) as model comparison fit indices. Table 5 summarized the reported model fit indices for each study and the criteria for the significance level.

Table 5

Summary of the Reported Model Fit Indices Criteria across the 10 Studies

Model Fit Indices	Study Assigned Number									
	1	2	3	4	5	6	7	8	9	10
0 ∈95% CI	X	X	X	X	X	X	X	X	X	X
PPP	≥ .05	>0	≥ .05	>0	>.001	≥ .05	≥ .05	≥ .05	> 0	≥ .05
DIC	-	-	X	-	-	X	X	-	-	-
BIC	-	-	-	-	-	-	-	-	-	-
Deviance	-	-	-	X	X	-	-	-	-	-
BF	-	-	-	-	-	-	-	-	-	-

Note. X= Study used the corresponding index; PPP= Posterior predictive *p*-value; 95% CI= 95% credibility interval of χ^2, BIC= Bayesian information criterion; DIC= Deviance information criterion; BF= Bayes factor.

Level of Invariance Achieved

Reported level of invariance before and after BAMI. All studies reported a level of MI before the BAMI. One study (10%; Muthén & Asparouhov, 2013) was not established any level of MI, and 20% of the studies (Braeken & Blömeke, 2016; de Bondt & van Petegem, 2015) held configural MI level. However, these two studies could not hold the configural level until the correlation errors or factors became a part of the model.

Although Muthén and Asparouhov (2013) insisted that the BAMI is efficient after the traditional full and partial MI failed, studies addressed this issue in various ways. All 10 studies emphasized that adopting the BAMI approach yielded advanced MI results compared to the traditional approach, (i.e., MGCFA). Reporting the level of invariance after BAMI indicated that the BAMI or partial BAMI is satisfied for all studies: 70% of the studies held BAMI scalar level,

10% of the studies held BAMI metric level, 10% of the studies held partial BAMI scalar and finally 10% of the studies held partial BAMI metric (see Table 6).

Table 6

Summary of Reported MI Level before BAMI and after BAMI

MI		Study Assigned Number									
		[a]1	2	3	4	[b]5	6	7	8	9	10
Before BAMI	Full	Config.		Metric		Config.	Metric	Metric	Metric		None
	Part		Metric		Metric					Metric	
After BAMI	Full	Scalar	Scalar	Metric	Scalar	Scalar	scalar	scalar	Scalar		
	Part									Scalar	Metric

Note. [a] this model used BSEM-based alignment with approximate measurement invariance. [b] approximate scalar with local item dependence. MI level= measurement invariance level holds; BAMI= Bayesian approximate measurement invariance; Full= full invariance; Part. = partial invariance; Config. = configural

Table 7

Number of Non-Invariant Loadings and Intercepts per Study before and after BAMI

Article Assigned Number	Scale			Number of non-invariant items	
	Factors	Items	Groups	Before BAMI	After BAMI
1	5	50	2	1 load.	0
2[a]	1	3	35 * 6	2-3 inter. within a wave.	0
3	2	9	2	4 inter.	4 inter.
4 [b]	19	48	8	9 factors inter. were non-invariance	0
5	1	8	12	2 inter.	0
6[c]	1	3	15 * 6=90	90-37=53 inter.	17 inter.
7	1	8	3	5 inter. in group 1, 4 inter. in group2	0
8	1	4	2	4 inter., Bias size (.193, .235, .167, .324)	0
9[d]	4	16	38	All (16) inter. for four scales	4 inter.
10	1	8	40	8 thresholds, 8 item difficulties	7 load., 25 inter.

Note. [a] study (2) has 6 waves, the number of groups which have bias items in each wave are: 2,4, 10, 10, 10, 10, 10; [b] study (4) has 19 factors, nine factors (values) showed non-invariance out of the 19 before BAMI; [c] Study (6) has 15 countries with 6 rounds =90 countries and count the item bias by country; [d] Study (9) has four scales, each scale has one factor and 4 items. BAMI holds for three scales and did not hold for the fourth one; BAMI= Bayesian approximate measurement invariant; load. = loading; inter.=intercept.

Number of non-invariant items. Before establishing the BAMI, all studies discussed the issue of items noninvariant, some per item and others per group. Similar discussions were made after the BAMI. Table 7 presented the reductions of the non-invariant items number after the application of BAMI. It is clear that all scales showed improvement in the invariance level. Six

studies showed full invariance, and four studies showed improved invariance levels, (i.e., better than their traditional MI levels). However, only the study of van de Schoot et al. (2013) provided the bias size per intercept, which ranged from .2 to .3.

Simulation Studies Review

Three BAMI simulation studies (Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013) were conducted to examine BAMI approach. They revealed a great promise of ascertaining measurement invariance of single level scales across many groups or over time. However, more studies are needed to establish a fundamental basis for BAMI approach and its model fit criteria. I will limit my review to discuss: 1) what were the simulation factors? 2) what Bayesian decisions they made including how the evaluation of model fit and model comparison was conducted? and 3) what were the challenges/ limitations they faced?

Table 8
Comparison of Bayesian Approximate Measurement Invariances Published Simulation Studies

	Muthén & Asparouhov (2013)	Van de Schoot et al. (2013)	Kim, Cao, Wang & Nguyen (2017)
Scale	Single factor /6 items	Single factor/ 4 items	Single factor/ 6 items
Groups Number	10	2	25, 50
Group Size	500	500	50, 100, 1,000
Number of Bias Items	4	2 and 4	2
Non-Invariant Parameter Location	Loading and intercept	Intercept	Intercept
Priors	.01, .05,.10	.005, .01, .05, .5	.001, .05
Percent of Groups with Non-Invariant Items	80%	50%	20% and 40%
Magnitude of Non-Invariance	Moderate (.2)	Small (.01) moderate (.1) large (.5)	Small (.0009) [a] large (.6)
Intercept Differences Direction	systematic	cancel each other	systematic
Model Fit and Model Comparison	PPP 95 % CI	PPP 95 % CI	PPP 95 % CI DIC BIC

Note. [a]This is not a prior variance. For the large DIF, they generated .6 difference in intercepts across groups.

The three BAMI simulation studies designed the simulation factors based on their purposes and objectives. Some simulation factors were shared across the three studies, (i.e., set up different priors values), and some were different, (i.e., sample size per group). A comparison among BAMI published studies using simulation conditions is shown in Table 8.

In the next section, a detailed discussion of the simulation factors across the three simulations studies, (i.e., Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013) is presented.

Review of Simulation Factors

Scale Length

Although the three simulation studies used different models, (e.g., Kim et al. (2017) and van de Schoot et al. (2013) used the CFA model whereas Muthén and Asparouhov (2013) used the IRT model), the length of the scale was similar across the three studies. Kim et al. (2017) and Muthén and Asparouhov (2013) used a single-factor scale with six items whereas van de Schoot et al. (2013) used a single-factor scale with four items.

Number of Groups and Group Size

Although the typical application of BAMI is the cross-cultural or cross-country research, the number of groups and group size were varied across the three simulation studies. Kim et al. (2017) used two numbers of groups (25 and 50) and three group sizes (50, 100, and 1,000). Van de Schoot et al. (2013) used one number of groups (2) and one group size (500). Finally, Muthén and Asparouhov (2013) used one number of groups (10) and one group size (500). Only balanced groups, (i.e., same sample size across groups) were used across the three studies.

Number of Biased Items

Out of a total of six items, Muthén and Asparouhov (2013) and Kim et al. (2017) used a fixed number of biased items: four items and two items respectively. Out of a total of four items, van de Schoot et al. (2013) used two numbers of biased items: two and four.

Non-Invariant Parameter Difference Location

Van de Schoot et al. (2013) and Kim et al. (2017) opted to locate the non-invariant differences in items intercepts only. Muthén and Asparouhov (2013) used two locations of non-invariant parameters differences: items loadings and intercepts.

Percent of Groups with Non-Invariant Items Intercepts

Because van de Schoot et al. (2013) had only two groups, 50% of the groups (i.e., one group) included the non-invariant items. Muthén and Asparouhov (2013) manipulated 80 % of the groups to include the non-invariant items. Finally, Kim et al. (2017) used two percentages of groups with non-invariant items: 20% and 40%. Under the 20% noninvariant groups, 5 out of 25 groups and 10 out of 50 groups had two noninvariant items. Under the 40% conditions, 10 out of 25 groups and 20 out of 50 groups had two noninvariant items.

Magnitude of Non-Invariance

Three sizes of non-invariance were observed across the three studies: small, moderate, and large. For small non-invariance size, (.0009) prior variance and (.1) were considered by Kim et al. (2017) and van de Schoot et al. (2013) respectively. Muthén and Asparouhov (2013) used only (.2) as a moderate size, whereas van de Schoot et al. (2013) used (.1) as moderate size. Finally, van de Schoot et al. (2013) used (.5), and Kim et al. (2017) used .6 difference as a significant magnitude of non-invariance.

Differences Direction

Kim et al. (2017) and Muthén and Asparouhov (2013) used a systematic direction for parameters differences (one direction), whereas van de Schoot used two parameters differences directions, (e.g., .01 versus -.01).

Review of Bayesian Decisions

BAMI Testing Approach

Kim et al. (2017) study tested BAMI in the intercept difference only by using two prior levels (.001) and (.05). A predetermined small-size prior variance of noninvariance (.001) was identified whereas a prior with a value of (.05) is considered substantial or large differences. Good fit of the model of (.001) or selection of this model over competing models (.05) is considered as the approximate scalar invariance. In other words, if the prior of (.001) was selected or the model fit showed the best model fit over other models, it indicated that approximate invariance held. On the other hand, if the prior of (.05) was selected, it indicated that approximate invariance did not hold. Van de Schoot et al. (2013) used four prior levels to test for BAMI: .5, .05, .01, and .005. Then, model fit results of each model were reported and evaluated. However, Muthén and Asparouhov (2013) used the two BAMI steps with three priors levels: .01, .05, and .10 (See BAMI testing procedure section earlier in this chapter).

Number of Replications

The execution time to run a Bayesian model is affected by the size of prior variance and the sample size. Therefore, a reasonable number of replications must be determined in advance. Both Kim et al. (2017) and Muthén and Asparouhov (2013) used 100 replications across conditions. Van de Schoot et al. (2013) used a large number of replications (i.e., 1,000) across conditions. However, their study has a limited number of groups, namely two.

Number of Iteration*s*

Another key decision in Bayesian simulation study is the number of sufficient iterations. Kim et al. (2013) study did not indicate the total number of iterations, but because M*plus* default was used, I inferred that they used 50,000. Van de Schoot et al. (2013) used 100,000 iterations with 5,000 as a minimum number of iterations; Muthén and Asparouhov (2013) used 10,000 as a minimum number of iterations without indication of the maximum. However, 50,000 of iterations is the default maximum number of iterations in M*plus*.

Number of MCMC Chains

Although Kim et al. (2013) did not mention the MCMC chain number, they indicated the use of M*plus* default, (i.e., 2 MCMC). The same number of chains was also used by Muthén and Asparouhov (2013). Finally, van de Schoot et al. (2013) used chain numbers: 4, 5, and 8.

Prior

Determining a prior value in BAMI is crucial because the accuracy of the results relies on using a suitable prior value. Different levels of priors precisions were used across the three studies, but a prior value of (.05) was common across all of them. Also, a prior value of (.01) was used in the study of van de Schoot et al. (2013) and the study of Muthén and Asparouhov (2013). Other prior values were used such as (.001), (.005), (.10), and (.5). Prior values of (.001), (.005), and (.01) were considered as small priors whereas (.05), (.10) and (.5) were considered as large priors.

Model Fit and Model Comparison Indices

The two Bayesian model fit indices (i.e., PPP and 95 % CI) were used in all of the three studies. The PPP value greater than .05 was considered as a good model fit whereas 95% CI

should include zero. For model comparison, only Kim et al. (2017) used BIC and DIC as indices to evaluate the best model.

Review of Challenges and Limitations

Variation across the three studies was not substantial. The similarities could be seen in the use of the same simulations factors such as the scale length, non-invariant parameter location, and items bias size. However, increasing the number of simulation conditions depends on the study design and purpose.

The numbers of replications and iterations were identical when comparing Kim et al. (2017) to Muthén and Asparouhov, (2013), whereas van de Schoot et al. (2013) utilized a large number of replications and iterations. Additionally, Muthén and Asparouhov (2013) used three prior values and Kim et al. (2017) used two prior values whereas van de Schoot et al. (2013) used four prior values.

Considering study purpose, the sample size across groups was sufficiently large, that is 500 per group in both studies of Muthén and Asparouhov (2013) and van de Schoot et al. (2013). However, Kim et al. (2017) used three sample sizes across groups because the primary purpose of Kim et al. (2017) study was to examine four MI approaches in addition to BAMI. Kim et al. (2017) also used two number of groups, (i.e., 25, 50), to fit well with all of the five MI approaches they studied.

Putting all together, the three BAMI simulation studies, (i.e., Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013) revealed a great promise of ascertaining approximate measurement invariance of single-level scales across many groups or over time. However, more studies are needed to establish a fundamental basis for BAMI approach and its model fit criteria.

Summary

In this chapter, challenges regarding the traditional approach of measurement invariance were addressed, and Bayesian approximate measurement invariance (BAMI) was introduced as a new level of measurement invariance testing. I investigated the BAMI approach regarding how it was implemented, and discussed the optimal BAMI usage, advantages and disadvantages, decisions within BAMI, and MI testing procedures. Then, a systematic review of 10 applied studies conducted between 2013 and 2017 was presented. The systematic review was conducted in terms of: a) the BAMI procedures followed; b) BAMI model evaluated; and c) level of invariance achieved. Finally, three BAMI simulation studies were examined, and their simulation factors and Bayesian decisions were presented. A brief discussion about challenges and limitations of these three studies was included. Results of this review stress the need to delineate guidelines to know how to utilize the BAMI estimation method, model fit evaluation and comparisons, and what to report in methodology and results sections.

Chapter Three: Method

In the previous two chapters, I discussed how the concept of measurement invariance can generally and operationally be applied across groups/times using the Bayesian approximate measurement invariance (BAMI). I reviewed (10) BAMI applied research studies and discussed their results. Additionally, I reviewed the three BAMI simulation studies, and compared their simulation factors and Bayesian decisions. However, more studies are needed to support their results and to add a fundamental basis for BAMI approach and its model fit criteria.

Simulation studies have known as an "excellent method for evaluating estimators and goodness-of-fit statistics under a variety of conditions, including sample size, nonnormality, dichotomous or ordinal variables, model complexity, and model specification" (Paxton, Curran, Bollen, Kirby, & Chen, 2001; p. 288). Therefore, the purpose of this dissertation was to extend aforementioned studies by examining how the BAMI model fit criteria behaved across different research settings. I extended previous research by evaluating the BAMI methodology under four conditions: a) number of groups (medium (8) and large (20)), b) percent of groups with non-invariant item intercepts (50% and 80%), c) the intercept differences directions (cancel each other out and systematic), and d) the magnitude of non-invariance (zero, small (.01), moderate (.2), and large (.6)). Further, in addition to the model fit criteria used in the previous simulation studies (PPP, 95% CI, BIC and DIC in Kim et al. (2017); PPP and 95 % CI in Muthén and Asparouhov (2013) and van de Schoot et al. (2013)), an investigation of the model fit comparison index (i.e., Bayes factor or BF) were conducted, which has not presented in any of

the BAMI applied or simulation studies. A summary of the comparison between the conditions used in previous BAMI simulation studies to those of the current research was shown in Table 9.

Table 9
Comparison of Published Simulation Studies on Bayesian Approximate Measurement Invariances to the Current Research

	Muthén & Asparouhov (2013)	van de Schoot et al. (2013)	Kim, Cao, Wang & Nguyen (2017)	The Current Study
Scale	Single factor /6 items	Single factor/ 4 items	Single factor/ 6 items	Single factor/ 6 items
Number of groups	10	2	25, 50	8, 20
Group size	500	500	50, 100, 1000	500
Bias item	4	2 and 4	2	4
Non-invariant parameter location	Loading and intercept	Intercept	Intercept	Intercept
Priors	.01, .05,.10	.005, .01, .05, .5	.001, .05	.001, .005, .01, .05, .10
Percent of non-invariant groups	80%	50%	20% and 40%	50% and 80%
Magnitude of non-invariance	Moderate (.2)	Small (.01) moderate (.1) large (.5)	Small (.0009) large (.6)[a]	Zero Small (.01) Moderate (.2) large (.6)
Intercept differences directions	systematic	cancel each other	systematic	cancel each other systematic
Replication	100	1000	100	100
Model fit and model comparison	PPP 95 % CI	PPP 95 % CI	PPP 95 % CI DIC BIC	Bayes factor PPP 95 % CI DIC BIC
Models	Exact invariance Bayesian Approximate	Scalar and Partial with (Exact-zero) Approximate-zero Partial approximate-zero	Scalar (Exact-zero) Scalar (Approximate-zero)	Scalar (Exact-zero) Scalar (Approximate-zero)

Note. [a]This is not a prior variance. For the large DIF, they generated .6 difference in intercepts across groups.

In the next section, I described the simulation design, data generation, fitting models, and dependent variables of the simulation study.

Simulation Design

I manipulated the population parameters in order to meet various simulation conditions. The BAMI method was carried out with one hundred replications for each condition. Details about the simulated data and factors were e presented in the next section.

Data Generation

The basic parameters were generated and the simulation factors were determined based on the results of the review of the BAMI applied and simulation research (see Chapter 2). Data were generated on the basis of a congeneric CFA model that has a single factor with six continuous items with homogeneous factor loadings under the assumption of multivariate normality, (i.e., metric invariance was met because the scalar level was only of interest). All items loaded on a single factor with the factor loadings of .8, .7, .6, .8, .7, .6 and intercept of zero. The residual variances of observed items were .36, .51, .64, .36, .51, and .64, respectively. These values were used in the previous BAMI simulation studies, (e.g., all loadings fixed at .80 and all residuals variances fixed at .36 by Kim et al., (2017); loadings as 1, .7, .5, 1, .7, and .5 and all residuals variances fixed at .5 by Muthén and Asparouhov (2013); loadings as .7, .6, .4, and .2 and fixed residuals at 1 by van de Schoot et al., (2013)).

The conditions that were manipulated in this study are: number of groups (medium (8), and large (20)), percent of groups with non-invariant items intercepts (50% and 80%), the intercept differences directions (cancel each other out and between groups and systematic), and magnitude of non-invariance (zero, small (.01), moderate (.2), and large (.6)). The total number of data generation conditions is 2*2*2*3= 24 for the conditions with DIF items and 2 for conditions without DIF items (exact zero).

Because typical applications for BAMI are country comparisons, I included a fixed group size with 500 observations per group considering large-scale data, such as international surveys with a large number of participants in each group (Muthén & Asparouhov, 2013). Also, previous simulation studies, (e.g., Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013), used group sizes of 500 and 1,000. The number of group factor (GN) based on 500 per group yielded the total sample size from 4,000 to 10,000. The simulation outcomes were evaluated based on the detection rates of the scalar noninvariance over the 100 replications. Detailed information can be found in the Simulation Outcomes section.

Type of Non-Invariance

The focus of this paper was the invariance of intercepts across groups (i.e., scalar level) for two reasons. First, scalar was the sufficient advanced MI level that must be achieved in order to conduct factor mean comparisons across groups (Brown, 2015; Davidov et al., 2015; Kim et al., 2017; Millsap, 2011; Vandenberg & Lance, 2000), which was the logic behind conducting MI in cross-cultural studies in most circumstances. Second, scalar invariance was the advanced MI level that was failed to hold most frequently when the traditional MI was used in a large number of groups (Desa, 2014; Nagengast & Marsh, 2013). Meanwhile, an improvement to the scalar invariance was reported by several studies when the BAMI approach was used, after the failure of achieving scalar level via traditional MI (e.g., Cieciuch et al., 2014; Davidov et al., 2015; He & Kubacka, 2015). Therefore, the scalar invariance level was the focus of this study. By assuming exact-zero metric invariance holds, I tested for scalar invariance (item intercept differences only).

Simulation Factors

Number of groups (GN = 8 and 20). The number of groups varies across studies based on research setting. For example, small number of groups was considered two or three groups (e.g., van de Schoot et al., 2013), medium number of groups was considered eight, twelve, and fifteen groups (e.g., de Bondt & Van Petegem, 2015; Zercher et al., 2015), and large number of groups was considered 23, 26, 38, and 40 groups (e.g., Beierlein, Davidov, Schmidt, Schwartz, & Rammstedt, 2012; He & Kubacka, 2015; Rutkowski & Svetina, 2014; Zercher et al., 2015). Similar numbers were also adopted in simulation studies such as 10, 20, 25, 30, and 60 (e.g., Kim et al., 2016, Kim et al., 2017; Muthén & Asparouhov, 2013). Therefore, I used GN= 8 and 20 because they were commonly observed in traditional MI and BAMI in applied studies. I did not include a small number of groups, such as two, since the practical application for BAMI approach was the large-scale assessment in cross-cultural research. Also, another MI approach, MGCFA, was known for reasonably detecting the MI in comparing two groups.

Percent of groups with non-invariant item intercepts (PCT= 50% and 80%). BAMI was better suited for a large number of minor non-invariant items (see Chapter 2; Muthén & Asparouhov, 2013). Previous BAMI simulation studies used balanced and unbalanced percent of non-invariant groups (groups with non-invariant item parameters). Van de Schoot et al. (2013) have only two groups, therefore, 50% of the groups (i.e., one group) included the non-invariant items. Muthén and Asparouhov (2013) have total of 10 groups where 80% of them were manipulated to include the non-invariant item parameters. Finally, Kim et al. (2017) have 20% and 40% groups with non-invariant items. Therefore, to better discover when BAMI performs well, I examined BAMI at two levels of groups with non-invariant item intercepts (50% and 80%) and the number of noninvariant items was fixed at four out of six items. I generated uniform non-invariance on four item intercepts in 50% and 80% of the large and medium

number of groups, which corresponds to 10 and 16 groups out of 20 when having a large number of groups and 4 and 6 groups out of 8 groups for the conditions of a medium number of groups.

With a medium number of groups (8) that had 50% of non-invariant groups, groups 1 to 4 had the biased items, and with 20 groups, groups 1-10 had the bias items. With a medium number of groups (8) that had 80% of non-invariant groups, groups 1 to 6 had the biased items, and with 20 groups, groups 1-16 had the bias items.

Magnitude of intercept differences (DIF-Size). The allowed magnitude of difference between items intercepts (wiggle room) was another simulation factor to consider. Through the BAMI literature, different intercepts differences levels were considered. In the published simulation studies, for examples, Kim et al (2017) considered .0009 variance as trivial intercept differences (and up to .6 higher intercept as large ones). Muthén and Asparouhov (2013) considered (.2 variance) as small differences whereas van de Schoot et al. (2013) used three levels of intercept differences: .01 variance as small intercept differences, .1 variance as moderated level, and .5 variance as large ones. In this study, I considered three levels of intercept differences (intercept variances): SM= small (.01), MD= medium (.2), and LG= large (.6). However, only the small intercept differences condition (.01) was considered as a representant of the approximate-zero size. In addition, zero noninvariance conditions were also simulated.

Direction of intercept differences (1 and 2). Muthén and Asparouhov (2013) stated that the ideal situation for using BAMI was a CFA model with a large number of items that have small parameter differences across groups, where these differences canceled each other out and between groups. Each of the three BAMI published simulation studies differently generated the direction of the intercept bias. Muthén and Asparouhov (2013) used a systematic direction of parameter differences in order to prove the failure of the BAMI approach within this context.

Van de Schoot et al. (2013) used items with small parameters differences across groups where these differences were canceled each other out. Kim et al. (2017) generated intercept differences systematically (in one direction). I created two directions in the intercepts differences across groups: (1) if the intercepts differences canceled each other out or (2) if these differences were systematic.

Table 10

Summary of Two Simulation Conditions: Magnitude and Direction of Intercept Noninvariance

Population	Number of Biased (DIF) Items	Intercept Differences Magnitude	Intercept Differences Direction
1 (exact)	0	Zero	-
2 (approximate)	4	Small	-.01 versus .01
3 (approximate)			.01
4 (non-invariance)		Medium	-.2 versus .2
5 (non-invariance)			.2
6 (non-invariance)		Large	-.6 versus .6
7 (non-invariance)			.6

For simplicity, I used DIF-Size= LG1, LG2, MD1, MD2, SM1 and SM2 to represent the two simulation conditions simultaneously, the magnitude and directions of item intercepts differences. Therefore, based on two simulation factors (magnitude and direction of intercept noninvariance), seven (7) populations were generated (see Table 10). The first one was the exact MI population (p #1) where all items loadings and intercepts were invariant. Six populations (p #2 - #7) had invariant item loadings but four items with differences in intercepts, items 1, 2, 3, and 4, where intercept differences are: (LG1) large and cancel each other (-.6 versus .6), and (LG2) large and systematic (.6), (MD1) medium and cancel each other (-.2 versus .2), (MD2) medium and systematic (.2), (SM1) small and cancel each other (-.01 versus .01), and (SM2)

small and systematic (.01). Only SM1, (small (.01) DIF in item intercepts that canceled each other) and SM2, (small (.01) systematic DIF in item intercepts) were considered as approximate-zero size. Large and medium DIF magnitudes were considered non-invariance size.

Prior Variance

Because the unique advantage of using Bayesian analysis was the researchers' abilities to propose their previous knowledge, testing different prior variances would be valuable to this research. Additionally, the choice of prior was essential in BAMI, where the definition of approximate or small differences in parameters across groups were not well established. I chose five different prior variance values: .001, .005, .01, .05, and .10. These priors were considered in several BAMI simulation and applied studies (see Table 4 in Chapter 2). Kim et al. (2017) used .001 and .05; Muthén and Asparouhov (2013) used .01, .05, and .1; and van de Schoot et al. (2013) used .005, .01, .05 and.5. As used in many studies, prior equaled to .01 or less (i.e., .001, .005) represented approximate-zero MI, and the prior equaled to .05 or greater (i.e., .10) represented the substantial non-invariance level.

Fitting Models

Two methods were considered in this study: the traditional exact-zero scalar MI test using maximum likelihood (ML) estimator and the approximate-zero scalar MI test using Bayes estimator. These two methods were used for all the populations (26 conditions) generated in this study.

Exact-zero scalar invariance test. Using the ML estimator, the exact-zero scalar invariance was tested with the generated data. To test exact-zero scalar MI, an exact-zero scalar invariance model that constrained both item loadings and intercepts equal across groups (zero

difference) was compared with an exact-zero metric invariance model where intercepts were allowed to be different across groups for all items except one reference item.

Approximate-zero scalar invariance (BAMI) test. Using this approach across all the populations, I allowed approximate-zero invariance in the intercept differences across groups. Five levels of precision for the priors were used (.001, .005, .01, .05, .10). For all other parameters, I used the default M*plus* priors settings (van de Schoot et al., 2013). To determine whether the Bayesian approximate-zero scalar invariance was held or not, I used two strategies that are found in the literature. First, the model with prior variance of .01 represented the approximate-zero scalar MI model (approximate scalar), whereas the model of prior variance of .05 represented the substantial (large) non-invariance (metric invariance). The prior variance of .01 (.1 SD) allowed wiggle room between -.2 and .2 (±2 SD) for item intercept differences between groups, which considered as approximate MI by Muthén and Asparouhov (2013). The prior variance of .05 (.22 SD) allows wiggle room between -.44 and .44 (±2 SD) for intercept differences between groups, which was considered as large non-invariance by Kim et al. (2017). For MI testing, the model with the prior variance .01 (namely, approximate-zero scalar invariance model) is compared to the model with the prior variance .05 (namely, substantial non-invariance model). The selection of the first model indicated the approximate-zero scalar invariance was held (Kim et al., 2017). Second, the models with five different priors were all compared, and the best fitting model was selected. When the prior variance was considered small (that is, .001, .005, and .01), I considered approximate-zero scalar invariance for this replication; when the prior variance was considered large (that is, .05, and .10), approximate-zero scalar invariance was rejected (Gucciardi et al., 2016; He & Kubacka, 2016; van de Schoot et al., 2013; Zercher et al., 2015). In the second approach, I also kept track of the prior selected as best fit

across simulation conditions to investigate which prior was the most optimal as a cutoff across conditions.

Estimation

In order to determine the convergence of the sampling procedure, several criteria for Bayesian estimation were applied. The posterior distribution of Bayesian estimation was achieved using the MCMC algorithm with the Gibbs sampler method. According to Muthén and Asparouhov (2012): "The idea behind MCMC is that the conditional distribution of one set of parameters given other sets can be used to make random draws of parameters values, ultimately resulting in an approximation of the joint distribution of all parameters" (p. 334).

To examine whether running the chain longer was necessary to identify local convergence problems and obtain a static statistic, a preliminary simulation study was conducted. The study's results showed that the MCMC samples were stable after a burn-in period between 20,000 and 30,000 iterations. Therefore, I determined that the sufficient number of iterations for the convergence would be around 50,000, (i.e., M*plus* default). In order to monitor the convergence, two MCMC chains with 50,000 iterations with the 25,000 burn-in period were specified. Different random seeds were used. The number of MCMC chains (2) was used in the BAMI studies, (e.g., de Bondt & van Petegem, 2015). This strategy was chosen based on previous Bayesian estimation studies (e.g., de Bondt & Van Petegem, 2015; Gucciardi & Zhang, 2016; van de Schoot et al., 2013). Gelman et al. (2014) stated that "posterior inferences concerning medians of posterior distributions are generally less sensitive to changes in the model than inferences about means." (p.185). Therefore, the medians of the posterior samples were taken after the MCMC procedure reached the maximum number of iterations.

Convergence Criteria

I assessed the MCMC convergence via the Gelman–Rubin convergence diagnostic, which used the potential scale reduction (PSR) factor with a PSR value below 1.1, which suggested model convergence (PSR < 1.1; Gelman et al., 2014; Gelman & Rubin, 1992). For each model, the PSR value indicated that the between-chain variation was small, relative to the within-chain variation. This must be reached before the first half (25,000) of the iterations were completed. The first half of the chains was a burn-in phase, and thus, it was discarded and the second half was used to estimate the posterior distribution (Muthén & Muthén, 2010). The PSR convergence criterion was used as M*plus* default when it printed "The Model Estimation Terminated Normally" across all replications.

Other criteria that were used to judge the convergence included visually checking the trace plots and the autocorrelation of the posterior distributions for model diagnostic, using M*plus* default thinning (Muthén, 2010). I checked the convergence visually across a few replications and then I relied on the numerical results because it was difficult and impractical to check all replications visually. Sampled parameter values over time were presented via trace plots where quick up-and-down variations and absence of long-term trends showed quick distribution convergence (de Bondt & van Petegem, 2015; Kaplan & Depaoli, 2012). The convergence diagnostic was used to compare the first and last halves of the post burn-in portion of the MCMC chain. In the MCMC chains, convergence occurred when the degree of correlation for parameter values across iterations (non-independence) was close to zero (0.1 or lower; Kaplan & Depaoli, 2012; Muthén, 2010). A further convergence check was done by looking for the results of the Kolmogorov-Smirnov test (k-s) and the improper prior statements under the "Technical 8" in M*plus* output. However, the k-s test was known to be too stringent whereas the

improper priors can still produce proper posteriors.[1] Finally, the convergence status and rate for each replication were recorded and reported across the simulation conditions.

Model Fit Evaluation

To evaluate the fit of models using ML estimator, goodness-of-fit indices were used based in Hu and Bentler (1999) cutoff criteria: chi-square test (χ^2) with degree of freedom (df = 1) and p value ≥ .05, the comparative fit index (CFI ≥ .95), the root mean square error of approximation (RMSEA ≤ .08), and the standardized root mean square residual (SRMR ≤ .08).

Bayesian models utilizing the approximate-zero invariance were evaluated with specific model evaluation strategies. Model fit was assessed via posterior predictive checking (Gelman et al., 1996) in order to test the structural model for misspecification. "The observed data should look plausible under the posterior predictive distribution." (Gelman et al., 2014, p. 143). So far, there was no clear-cut PPP value to indicate whether or not model fit was acceptable, but when a model was misspecified, the PPP was expected to be extreme (Kim et al., 2017). PPP was interpreted as goodness-of-fit indices in a structural equation modeling where a bigger PPP close to .5 indicated a better model. Low PPP values close to zero (<.01) or high PPP values close to 1 (>.95) indicated poor model fit (Gelman et al., 2014; Muthén & Asparouhov, 2012). Gelman et al. (2014) stated that "if a p-value is close to 0 or 1, it is not so important exactly how extreme it is." (p.150). Therefore, a PPP value between 0.05 and 0.95 was considered reasonable (Gelman et al., 2014). Additionally, the 95% CI for the difference between observed and replicated chi-square values were used, and it should include zero (Gelman et al., 2014; Muthén &

[1] In M*plus* online discussion under structure equation modeling, Bayesian BSEM structural invariance, Muthén (2015) stated that both the K-S test and the improper prior statement could be ignored since they found "the K-S test to be too strict and improper priors can still lead to proper posteriors which is all that matters". For more information check Mplus discussion at **http://www.statmodel.com/discussion/messages/11/12237.html?1485882536**

Asparouhov, 2013). If the 95% CI did not include zero, it indicated model misfit. Positive 95% lower limit suggested a poor model fit. For an excellent model fit, a posterior predictive p value should be around .5 and a symmetric credibility interval should be centering close to zero.

For model comparisons using ML estimator, the likelihood ratio test (LRT), the Chen (2007) and Cheung and Rensvold (2002) criteria, Bayesian information criterion (BIC) and Akaike information criterion (AIC) were used: a more constrained model with invariance constraints is selected if LRT $p \geq .05$; change in the comparative fit index (Δ CFI) $\leq .01$; change in the root mean square error of approximation (Δ RMSEA) $\leq .015$; BIC and AIC are smaller.

For model comparisons using Bayes estimator, three model comparison indices were used: BF, BIC, and DIC. Generally speaking, a model with a smaller value of DIC and/or BIC was preferred (Kaplan & Depaoli, 2012; Kass &Raftery, 1995). For BF, according to Kass and Raftery's (1995) rule of thumb, a BF value between 1 to 3 indicated weak support, 3 to 20 indicated positive support, 20 to 150 indicated strong support, and BF >150 indicated very strong support.

Analyses Procedures

BAMI testing can be conducted via two-step analysis process: 1) researchers specified BSEM by replacement of parameter specifications for exact-zeros differences with approximate-zeros based on informative small-variance, and then, 2) freeing the non-invariant parameters. In order to use the first step only, Muthén and Asparouhov's (2013) recommended for future BAMI researchers to generate data "with many non-invariant parameters and where the non-invariance is in both direction and more in line with BSEM specification" (p.18). They stated that there will be no need for freeing the non-invariant parameters because the first step is sufficient. Therefore, I used the first step of BAMI because the data were generated in line with BSEM specification.

In this study, two methods of MI testing were considered: exact-zero scalar MI testing (with ML estimator) and approximate-zero MI testing (with Bayes estimator). For the first test, (i.e., exact-zero scalar invariance), identical loadings and intercepts were specified across groups. The generated data were fitted on the exact-zero scalar invariance model. The data were also fitted to the exact-zero metric invariance model. Models of exact-zero metric invariance and exact-zero scalar invariance were evaluated with goodness-of-fit indices such as χ^2, CFI, RMSEA, and SRMR based in Hu and Bentler (1999) cutoff criteria. Model comparisons between exact-zero metric invariance and exact-zero scalar invariance models were conducted using the LRT, Δ CFI, Δ RMSEA. I also used BIC, and AIC for model comparison.

On the second method, (i.e., Bayesian approximate-zero scalar invariance testing), I tested approximate scalar invariance across populations 2 through 6 in Table 10 (that is, small, medium, and large noninvariance conditions excluding the exact-zero or no DIF conditions). I allowed for approximate-zero variance in intercepts differences by specifying five levels of precision for priors (.001, .005, .01, .05, .10). For all other parameters, I used the default M*plus* prior settings (van de Schoot et al., 2013). I identified the Bayesian approximate-zero invariance prior variance for discrepancies in intercepts as .01 because intercepts difference lay between ± .2, based on the applied study by Bujacz et al. (2014) and the simulation study by van de Schoot et al. (2013). I also identified the substantial non-invariance prior variance for discrepancies in intercepts as .05 because intercepts difference lay between ± .44, based on applied study by Muthén and Asparouhov (2013) and the simulation study by Kim et al. (2017).

A series of Bayesian approximate-zero scalar MI models with several prior variances, .005, .001, and .10 in addition to the predetermined priors (.01 and .05), were performed. Models were evaluated with two Bayesian fit indices: PPP > .05 and 95% CI should include zero

(Muthén & Asparouhov, 2013; Gelman et al.2014; van de Schoot et. al., 2013). According to Muthén and Asparouhov (2013), "if the prior variance is small relative to the magnitude of non-invariance, PPP will be lower than the prior variance corresponds better to the magnitude of non-invariance" (p.21).

A model comparison was conducted, and the best model fit was selected by BIC and DIC (i.e., smaller value as indicative of a better model). For BF, which was never used in the BAMI applied or simulation research, I investigated different cutoff points, (i.e., BF > 3 and BF > 20). If these two cutoffs did not work, I considered 150. These values were suggested in the literature by Kaplan and Depaoli (2012), Kass and Raftery (1995), Muthén and Asparouhov (2012b), and Verhagen et al. (2016).

In order to achieve the approximate-zero scalar invariance level, the prior variance in the selected model is supposed to be smaller or equal to the approximate-zero invariant prior variance for intercept differences that was determined in advance (.01) (analogous to supporting approximate-zero scalar invariance). If the selected model had a prior equal or greater than (.05), I rejected approximate-zero scalar MI (analogous to supporting non-invariance model) because the intercept differences were considered substantial.

Simulation Outcomes

Prior to analyzing any results, nonconvergent solutions were screened, and discarded from the analysis. Replications with convergence satisfaction out of 100 were checked across conditions. Lack of convergence may occur due to several reasons, such as poorly specified model, starting values, or lack of identification (Bandalos & Gagne, 2012). Therefore, the degree to which the nonconvergence occurred was significant to understand the impact of the simulation

factors. The convergence rate for Bayesian estimation, which was the proportion of replications in which estimation reached the convergence, was recorded and summarized.

To answer the first research question, for each model, I examined which level of MI was detected using selection criteria to measure the proportion of detecting scalar noninvariance, (i.e., rejection of scalar invariance). In case of the exact zero (no DIF) populations, the detection rates of LRT are in fact Type I error rates. The detection rates were summarized by number of groups, percent of groups with non-invariant item intercepts, and direction and size of differences of the non-invariant item intercepts.

For each replication from the Bayesian models, the behaviors of the Bayes factor (BF), deviance information criterion (DIC), and the Bayesian information criterion (BIC) in rejecting approximate scalar invariance were counted. The detection rate of scalar noninvariance was computed across the 100 replications. Table 11 showed which model was the generated model and which model was the corresponding correctly-specified model for the seven populations under ML exact-zero scalar invariance testing and under Bayesian approximate-zero scalar invariance testing.

To answer the second research question, the impact of each of the simulation design factors (i.e., group numbers, percent of groups with non-invariant items intercepts, and direction and magnitude of non-invariance) on detecting scalar noninvariance were examined. A description of the impact of each simulation factor on the BAMI was provided.

Additionally, the prior precision values that were selected as the best fitting model were collected and summarized across simulation conditions. This summary provided insights about a reasonable cutoff of the prior variance that could be used for Bayesian approximate-zero scalar MI testing in different research settings.

Table 11

Summary of the Generated Population with Its Corresponding Correctly-Specified invariance Model

Population	Number of Biased Items	Intercept Differences Size and Direction	First Model: Exact-Zero Scalar Invariance		Second Model: Bayesian Approximate-Zero Scalar Invariance	
			Scalar Invariance	Metric Invariance	Approximate Scalar Invariance Using Priors (.001, .005,.01)	Non-Invariance Using Priors (.05,.10)
1 (exact)	0	0	X			
2 (approximate)	4	-.01 versus .01		X	X	
3 (approximate)		.01		X	X	
4 (non-invariance)		-.2 versus .2		X		X
5 (non-invariance)		.2		X		X
6 (non-invariance)		-.6 versus .6		X		X
7 (non-invariance)		.6		X		X

Note. X represents the correctly specified model corresponding to the generated population

Summary

The purpose of this dissertation research was to build on previous work by exploring and learning more about how the Bayesian approximate measurement invariance model fit criteria behaved across different research settings. I extended the previous research by evaluating the BAMI methodology under four conditions: a) number of groups (medium (8), and large (20)), b) percent of groups with non-invariant items intercepts (50 % and 80%), c) the directions of the intercept differences (canceling each other out both and between groups and systematic), and d) the magnitude of non-invariance (small (.01), moderate (.2), and large (.6). Further, five levels of prior estimates (.001, .005, .01, .05, .10) were used. In addition to the model fit criteria that were used in Muthén and Asparouhov (2013) and van de Schoot et al. (2013), (i.e., PPP and 95 % CI), and Kim et al. (2017) PPP, 95 % CI, BIC and DIC, an investigation of a fit comparison index, (i.

e., BF), was conducted, which was not presented in any of the BAMI applied or simulation research.

Chapter Four: Results

Chapter 4 provides the results of the current simulation study. This study intended to investigate the behavior of the Bayesian approximate measurement invariance (BAMI), scalar level in particular, under different design factors. The simulation factors include number of groups, percent of groups with non-invariant item intercept, the intercept differences directions, and magnitude through addressing the following questions:

1) What is the performance of the model fit criteria on the BAMI testing in detecting non-invariant level across groups in the single level CFA?
2) What impacts do the design factors (i.e., number of groups, percent of groups with non-invariant item intercepts, the direction and magnitude of non-invariant item intercepts) have on the simulation outcomes of testing and estimating the approximate measurement invariance?

Twenty-six conditions under seven population scenarios, each with 100 replications, were generated. Two types of measurement invariance (MI) testing were conducted: 1) Exact-zero invariance testing using maximum likelihood (ML) estimator and 2) approximate-zero invariance testing with five prior levels .001, .005, .01, .05, and .10 using Bayes estimator. Two types of analyses were used to evaluate the performance of these methods. First, convergence rates and model fit evaluation for models using ML and Bayes estimators were provided. Second, detection rates for measurement invariance (exact-zero scalar, approximate-zero scalar) models were evaluated and interpreted. Additionally, the prior precision values that are selected as the

best fitting model are collected and summarized across simulation conditions in order to provide a reasonable cutoff of the prior variance. Because the Bayesian approach conceptually and methodologically differs from likelihood-based estimation methods, each estimator results are presented separately. Across this chapter, abbreviations were used interchangeably for the design factors, (i.e., number of groups (GN), percent of groups with non-invariant item intercepts (PCT)). Also, the term DIF was used interchangeably with DIF-Size to represent the non-invariance or the magnitude of difference in item intercepts. Finally, as another way to represent the PCT condition in an understandable way, the terms "balanced and unbalanced groups" were used and defined as equal (50%) and unequal (80%) percent of groups that including non-invariant item intercepts respectively.

Models Estimations Convergence Rates

Exact-Zero Invariance Testing

All the seven populations across all conditions were fitted into exact-zero metric and exact-zero scalar invariance models using ML estimator. The convergence rates for both models were computed across replications. Both exact-zero metric and scalar models were 100% converged across replications and simulation conditions, (i.e., the proportion of inadmissible solutions equaled to .00).

Approximate-Zero Invariance Testing

Similar to the previous step, except the exact population that has zero differences in item loadings and intercepts across groups, six populations were fitted into BAMI models with different levels of priors. The MCMC convergence was assessed using the potential scale reduction (PSR) factor with a PSR value below 1.1, as M*plus* default when it printed "The Model Estimation Terminated Normally" across all replications. (see Chapter 3 under Convergence

Criteria). The convergence rates were observed across all the simulation conditions and yielded 100%.

Further plots and diagnostics were used as other criteria to judge the convergence of proportion of the replications. Visual inspections of the trace plots and the autocorrelation across a random sample of replications, ranged between 5 and 20 replications for each condition under both models, were examined.

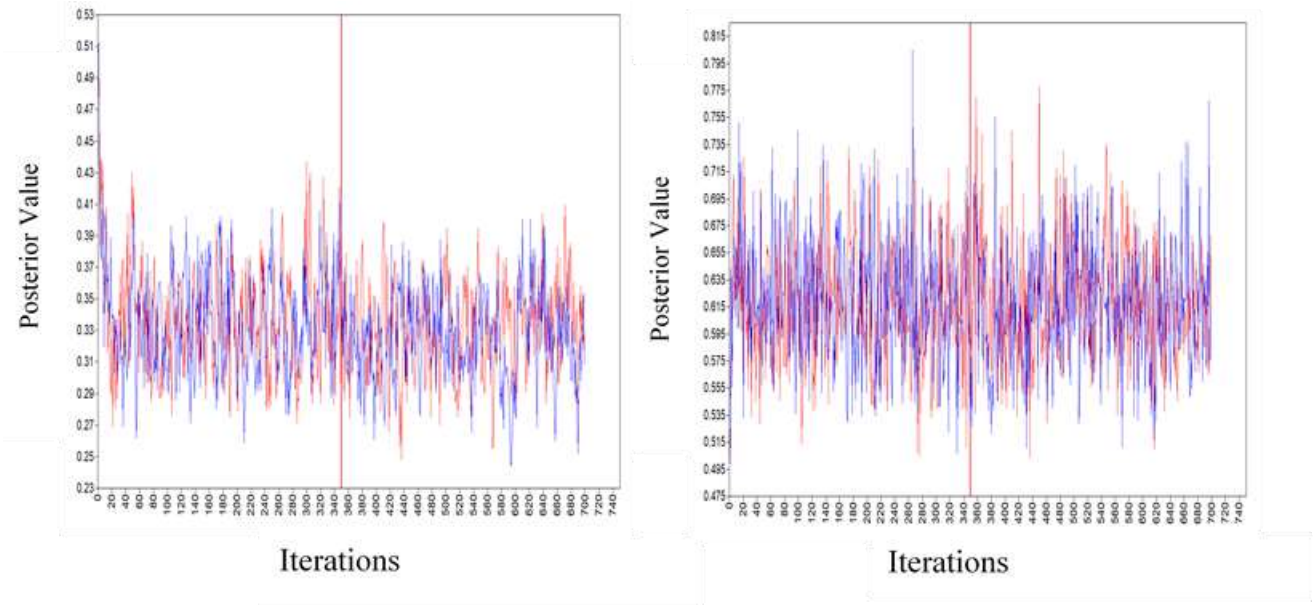


Figure 4. Random sample of trace plots to judge the convergence.
Only the last 25.000 (after the red vertical line) are used for the parameter estimates.

The trace plots looked mostly good with quick up-and-down variations and absence of long-term trends. Figure 4 showed a sample of the random trace plots that showed quick up-and-down variations and absence of long-term trends. To rest assured, a further check of parameter estimates of selected replications (especially when the trace plots were not ideal) showed that the parameters recovered correctly.

Also, the sample of random autocorrection plots showed a drop with increasing *k* (or lag the x-axis in the plot), which was a good sign. Figure 5 showed the autocorrelations between the samples and lag, (lag-*k* autocorrelation), the correlation between every sample and the sample *k* steps before), that returned by the Markov Chain Monte Carlo Chains (MCMC).

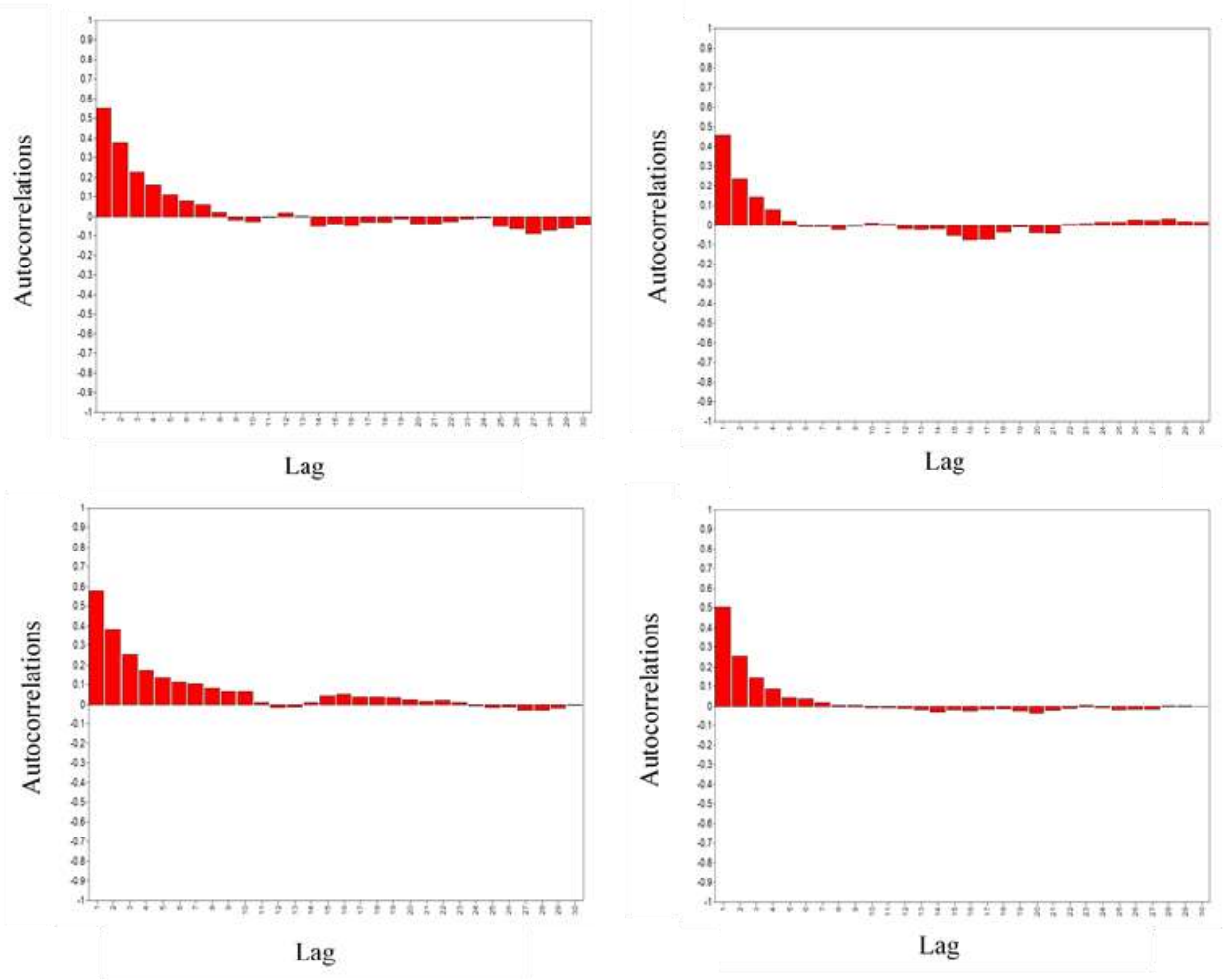


Figure 5. Random sample of autocorrelations plots between the samples returned by the Markov Chain Monte Carlo Chain (MCMC).

Finally, in a random sample of replications and under "Technical 8" in M*plus* output, the Kolmogorov-Smirnov test (k-s) and the improper prior statements were checked. Across the random sample of replications, k-s test results mostly showed nonsignificant *p*-values, except

one or two parameters that were differed across the random replications. Most of the parameters showed a good K-S test results with no improper prior statements, Therefore, I trusted the numerical results (PSR) as a quantifiable measure of convergence because it was difficult and impractical to check all replications. I concluded that the convergence was reached.

Model Fit Assessment

Exact-Zero MI Test with ML Estimation

The exact-zero metric invariance model (i.e., no difference in item loadings only across groups) was fitted into all seven populations with exact, LG1, LG2, MD1, MD2, SM1 and SM2 across simulation conditions. Results of goodness-of-fit indices (i.e., χ^2, CFI, RMSEA, and SRMR) showed that exact-zero metric invariance has excellent model fit, based in Hu and Bentler (1999) cutoff criteria, across all the seven populations either with a large or small number of balanced and unbalanced groups (see Chapter 3).

The same scenario was repeated to test the seven populations for the exact-zero scalar invariance, (i.e., zero differences in item loadings and intercepts across groups). As expected, the scalar invariance models under exact and small DIF magnitude (or approximate invariance) conditions produced excellent model fits. In contrast, conditions DIF-Size= LG1 and LG2, models showed poor fit regardless of GN and PCT. Although models with DIF-Size= MD1 and MD2 showed good model fit across all conditions, models with DIF-Size= exact, SM1 and SM2 fitted the scalar model better. Therefore, it was evidenced that the size of item intercept differences (DIF-Size) in the non-invariant items was generated successfully across conditions. Because results are similar when having either 8 or 20 groups, Table 12 showed only the results of scalar invariance model across all populations when having 20 groups with exact, 50%, and 80% of groups with non-invariant item intercepts.

Table 12

Summary of Means of Goodness-of-Fit Indices after Applying the Exact-Zero Scalar Invariance across Simulation Conditions.

GN	PCT	DIF-Size	χ^2	*df*	*P*	RMSEA	CFI	SRMR
	0	**Exact**	**390.74**	**370**	**0.304**	**0.01**	**0.998**	**0.032**
	50	LG1	3552.71	370	0	0.131	0.809	0.095
		LG2	5184.64	370	0	0.161	0.71	0.128
		MD1	765.57	370	0	0.046	0.976	0.043
		MD2	961.03	370	0	0.056	0.964	0.051
		SM1	**389.45**	**370**	**0.333**	**0.009**	**0.999**	**0.032**
20		**SM2**	**388.30**	**370**	**0.317**	**0.009**	**0.999**	**0.032**
	80	LG1	2378.76	370	0	0.104	0.879	0.078
		LG2	3234.90	370	0	0.124	0.827	0.105
		MD1	636.24	370	0	0.038	0.984	0.04
		MD2	767.01	370	0	0.046	0.976	0.045
		SM1	**387.81**	**370**	**0.337**	**0.009**	**0.999**	**0.032**
		SM2	**398.57**	**370**	**0.243**	**0.011**	**0.998**	**0.032**

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; Exact= zero differences in item loadings and intercepts across groups; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; χ^2= chi-square test, *df*= degree of freedom; p ≥ .05; CFI ≥ .95 the comparative fit index; RMSEA ≤ .08 the root mean square error of approximation; SRMR ≤ .08 the standardized root mean square residual; **bold**= good fit results.

Approximate-Zero MI Test with Bayes Estimation

Two Bayesian fit criteria, (i.e., 95% credible interval (95% CI) and posterior productive p-value (PPP)), were used to assess the approximate-zero scalar invariance model with five prior precision levels, (i.e., .001, .005, .01, .05, and .10). Models with PPP value larger than 0.05 and 95% CI including zero were considered as having a reasonable model fit. Table 13 presents the proportions of good model fit (95% CI including zero; ppp > .05) when the Bayesian approximate-zero measurement invariance model was fitted with five levels of priors.

Table 13

Summary of the Proportion of Good Fit of Bayesian Approximate-Zero Scalar Invariance Models with all Five Priors across the Simulation Conditions

GN	PCT	DIF-Size	Prior									
			.001		.005		.01		.05		.10	
			CI	PPP	CI	PPP	CI	PPP	CI	PPP	CI	PPP
20	50	LG1	0	0	0	0	.11	.04	.99	.98	.99	.98
		LG2	0	0	0	0	0	0	.99	.99	1	.99
		MD1	0	0	.96	.93	1	1	1	1	1	1
		MD2	0	0	.70	.61	.97	.96	1	.98	1	.98
		SM1	.99	.99	1	1	1	1	.99	.99	.99	.99
		SM2	1	1	1	1	1	1	1	.99	1	.99
	80	LG1	0	0	0	0	.51	.36	.98	.97	.99	.97
		LG2	0	0	0	0	0	0	1	1	1	1
		MD1	.08	.03	.99	.99	1	1	1	1	1	1
		MD2	0	0	.91	.86	1	1	1	1	1	1
		SM1	1	1	1	1	1	1	1	1	1	1
		SM2	1	1	1	1	1	1	1	.99	1	.99
8	50	LG1	0	0	0	0	.46	.30	1	1	1	1
		LG2	0	0	0	0	0	0	1	1	1	1
		MD1	.10	.07	.99	.97	1	.99	1	1	1	1
		MD2	0	0	.92	.89	1	1	1	1	1	1
		SM1	1	.99	1	1	1	1	1	.99	1	.99
		SM2	1	1	1	1	1	1	1	1	1	1
	80	LG1	0	0	0	0	.71	.66	.99	.99	.99	.99
		LG2	0	0	0	0	.17	.11	1	.99	1	1
		MD1	.37	.32	1	1	1	1	1	1	1	1
		MD2	.03	0	.97	.95	1	1	1	1	1	1
		SM1	1	1	1	1	1	1	1	.99	1	.99
		SM2	1	1	1	1	1	1	1	1	1	1

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; CI= 95% credible interval included zero; PPP= Posterior productive *p*-value > .05.

The model fit results for all priors showed a variation in fit quality based on the direction and the size of the associated DIF (DIF-Size). As expected, small prior variances showed reasonable fit with small DIF conditions but poor fit with medium and large DIF conditions in general. Models with .001 prior fitted both small DIF-Sizes very well but model fit deteriorated with other DIF-Sizes, (i.e., LG1, LG2, MD1, MD2). Also, models with .005 and .01 prior variances fitted all DIF-Size models well except models that are associated with large DIF-Size conditions (LG1 and LG2). Of note, when the prior size was large, it fitted well across all conditions regardless of the DIF magnitude or direction, and thus, models that have .05 and .10

prior variances almost always showed good fit. The prior .001 was sensitive to moderate and large size of DIF; the priors .005 and .01 were sensitive to large DIF; the priors .05 and .10 were insensitive to any size of DIF generated in this study. These results were partly supported and found in previous Bayesian simulation and applied studies because a prior variance that equaled to .01 or less, (i.e., .001, .005) was a representant of approximate-zero scalar MI whereas prior variance that equaled to .05 or greater (i.e., .10) was a representant of the substantial non-invariance level (see Chapter 2).

The Detection Rates

Throughout the paper, the detection rate was defined as the proportion of replications in which DIF was detected or scalar invariance was rejected. When testing exact-zero scalar invariance with likelihood ratio tests (LRT; a significance level of 0.05) for the exact populations (no DIF), the detection rates are in fact Type I error rates of falsely detecting DIF because exact-zero scalar invariance were true in the population and Type I error is the probability of rejecting the correct null hypotheses (i.e., scalar invariance). For the conditions of other DIF magnitudes, the detection rate was the proportion of replications in which the scalar invariance was rejected. For models using Bayes estimator, the detection rates are also provided when testing the scalar BAMI, (i.e., the proportion that a larger prior variance model supporting DIF was selected against a smaller prior variance model when two models of different priors were compared). The detection rates are summarized by the four simulation conditions: number of groups, percent of groups with non-invariant item intercepts, and the direction and the magnitude of differences of the non-invariant item intercepts as seen in Tables 14 through 20.

Detection Rates of Exact-Zero Scalar MI Models Using ML Estimator

The detection rate was computed for LRT, ΔCFI, ΔRMSEA, BIC, and AIC as the following descriptions. The measurement invariance was rejected or DIF was detected when LRT p value < .05; Δ CFI > .01 (Δ CFI = $CFI_{metric} - CFI_{scalar}$); Δ RMSEA > .015 (Δ RMSEA = $RMSEA_{scalar} - RMSEA_{metric}$). For BIC and AIC, a model with a smaller value was selected.

Exact-Zero Scalar Invariance Test with the Exact Population

In general, with the exact-zero scalar invariance, Type I error was inflated above .05 when using LRT, but Δ CFI, Δ RMSEA, BIC and AIC mostly supported the exact-zero scalar invariance.

Table 14

Type I Error Rates of Fitting the Exact-Zero Scalar Invariance to Exact Population

GN	PCT	DIF-Size	LRT	Δ CFI	Δ RMSEA	BIC	AIC
20	0	Exact	.36	0	0	0	0
8	0	Exact	.21	0	.05	0	.02

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; Exact= zero-DIF in items loadings and intercepts; LRT= the likelihood ratio test and $p \geq .05$; the change in the comparative fit index Δ CFI ≤ -.01; the change in the root mean square error of approximation Δ RMSEA ≤ .015; the smaller the better for BIC=Bayesian information criterion and AIC= Akaike information criterion.

Table 14 presents the exact-zero scalar model fitted to the exact population on 20 and 8 groups. As seen, the rates of Type I error under LRT were .36 and .21 with a large and medium number of groups respectively. As expected, Type I error rates were higher with the larger sample size. Therefore, detection rates of ΔCFI, ΔRMSEA, BIC, and AIC were examined. The ΔCFI 100% supported scalar invariance for both 20 and 8 groups. The same was true when ΔRMSEA with 20 groups (100% support scalar) and with 8 groups 95% supported scalar invariance. BIC and AIC also supported the exact scalar invariance model with 0% DIF detection rates when having 20 groups and with 0% and 2% when having 8 groups. These results

were expected because the exact-zero scalar invariance model fitted the exact population when using χ^2, CFI, RMSEA, and SRMR as model fit criteria.

Exact-Zero Scalar Invariance Tests with The Non-Invariance Populations

Applying the scalar invariance tests into different DIF-Sizes, (i.e., SM1, SM2, MD1, MD2, LG1, and LG2) produced variations in the detection rates. Table 15 presents the detection rates across conditions.

Table 15
Detection Rates of Testing Exact-Zero Scalar Invariance for Non-Invariance Populations

GN	PCT	DIF-Size	LRT	Δ CFI	Δ RMSEA	BIC	AIC
20	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	1
		MD2	1	1	1	1	1
		SM1	.35	0	.01	0	0
		SM2	.35	0	0	0	0
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	.99	1	0	1
		MD2	1	1	1	0	1
		SM1	.37	0	0	0	0
		SM2	.37	.01	.04	0	0
8	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	1
		MD2	1	1	1	.23	1
		SM1	.16	0	0	0	0
		SM2	.24	0	.04	0	.03
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	.95	1	1	1
		MD2	1	1	1	0	1
		SM1	.28	0	.03	0	0
		SM2	.25	0	.03	0	.02

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; LRT= the likelihood ratio test and p ≥ .05; the change in the comparative fit index Δ CFI ≤ -.01; the change in the root mean square error of approximation Δ RMSEA ≤ .015; the smaller the better for BIC=Bayesian information criterion and AIC= Akaike information criterion.

With large DIF conditions, LRT, ΔCFI, ΔRMSEA, BIC, and AIC 100% favored the exact-zero metric invariance against the exact-zero scalar invariance across all simulation conditions. For the medium DIF magnitude (MD1 and MD2), high detection rates, 1 or close to

1, were found to support the exact-metric invariance as well with all tested statistics except BIC. When approximate-zero scalar invariance was generated in the population, (SM1 and SM2), the detection rates that detected DIF and rejected scalar invariance were found to be less than 5% except LRT. Based on the model fit assessment, scalar invariance showed excellent model fit when fitted into populations with small DIF, in addition to the exact population, when using χ^2, CFI, RMSEA, and SRMR as model fit criteria. In contrast, populations with medium and large DIF conditions showed poor model fit when testing for exact-zero scalar invariance.

Detection Rates of Approximate-Zero Scalar Invariance Models Using Bayes Estimator (BAMI)

The proportion of replications in which the DIF was detected (or approximate-zero scalar invariance was rejected), that is, the model with a larger prior variance was selected, was computed for Bayes factor (BF), BIC, and DIC as the following descriptions. Because this is an exploratory study in terms of using BF in BAMI, three cutoff points from the literature were examined (i.e., BF ≥ 3, BF ≥ 20, and BF ≥ 150). When models with two different prior variances (e.g., .01 vs. .05) were compared, BF was the ratio of the model with a larger prior to smaller. When the cutoff was 3, if BF ≥ 3, the model with a larger prior was selected, that is, approximate scalar invariance was rejected. For the cutoff 20, if BF ≥ 20, scalar invariance was rejected. For the cutoff 150, if BF ≥ 150, scalar invariance was rejected. For BIC and DIC, when the DIC and BIC supported a model with a larger prior variance (e.g., .05 instead of .01) with a smaller value, DIF was detected and scalar invariance was rejected.

Detection Rates of Bayesian Approximate-Zero Scalar Invariance (BAMI) Testing When Comparing .05 Prior Model against .01 Prior Model Using BIC, DIC and BF

Table 16 presented the detection rates obtained from running the BAMI tests using two prior precision models: with .01 and .05 prior variances. The reported detection rates were the proportions that the .05 prior model was selected over .01 prior model for all relevant simulation conditions. The detection rates were expected to be very minimal and close to zero for small DIF-Size conditions, (i.e., SM1, SM2; approximate-zero invariance conditions) as favoring .01 prior models, whereas high and close to 1 for moderate and large DIF-Size conditions, (i.e., MD1, MD2, LG1, LG2; non-invariance conditions) in supporting .05 prior models.

BIC. BIC did not produce reliable results when models of .01 and .05 priors were compared. Across all conditions regardless of DIF magnitude, BIC 100% selected a model with the bigger prior .05 over a model with .01 (rejected the scalar invariance) and failed to differentiate approximate scalar MI. These results aligned with previous research that studied and described the behavior of BIC, (e.g., Kim et al., 2017).

DIC. The behavior of the DIC in detecting the .05 prior models was promising and consistent with previous studies (Kim et al., 2017) except for the medium DIF-Size conditions (MD1 and MD2). The DIC performed very well with high detection rates (1) for large DIF conditions and the low detection rates (< .06) for small DIF (approximate-zero invariance) conditions. However, the medium DIF magnitude conditions have generally low detection rates. Especially when the differences canceled out each other (MD1), the detection rates ranged between .03 and .31 in comparison to the range .31 and .94 for MD2. The low detection rates were more serious with unbalanced groups (e.g., .03 when GN=20 and DIF-Size=80%).

Bayes factor (BF). First, the BF_3 detected moderate and large DIF with 100% detection rates. However, the BF_3 frequently rejected scalar invariance when approximate scalar

invariance was simulated (i.e., SM1 and SM2). The detection rates under these conditions ranged between .74 and 1.

Then, BF_20 highly detected the large DIF, LG1 and LG2, with 100% detection rates and moderately detected the medium DIF with detection rates that ranged between .72 and 1. Also, BF_20 sometimes detected the small DIF with medium group numbers (e.g., .08 to .13 detection rates under GN=8, DIF-Size=SM1 and SM2). Unfortunately, with large group numbers, that is GN=20, the BF_20 often rejected scalar invariance when approximate-zero scalar invariance was simulated (i.e., SM1 and SM2) with detection rates that ranged between .49 and.53.

Table 16

Detection Rates of Bayesian Approximate-Zero Scalar Invariance Tests When Comparing .05 Prior Model against .01 Prior Model

Conditions			Model Comparisons Criteria				
GN	PCT	DIF-Size	BF_3 Prior .05	BF_20 Prior .05	BF_150 Prior .05	BIC Prior .05	DIC Prior .05
20	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	.14
		MD2	1	1	1	1	.94
		SM1	1	.49	.15	1	0
		SM2	1	.52	.04	1	0
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	.98	1	.03
		MD2	1	1	1	1	.31
		SM1	1	.53	.08	1	0
		SM2	.99	.50	.04	1	0
8	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	.89	.30	1	.28
		MD2	1	1	.97	1	.77
		SM1	.83	.08	.02	1	.06
		SM2	.85	.13	.02	1	.06
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	.72	.33	1	.31
		MD2	1	1	.72	1	.59
		SM1	.79	.13	.01	1	.04
		SM2	.74	.09	.03	1	.05

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; BF_3= Bayes Factor using BF ≥ 3 as a cutoff; BF_20= Bayes factor using BF ≥ 20 as a cutoff; BF_150= Bayes Factor using BF ≥ 150 as a cutoff; BIC=Bayesian information criterion; DIC= Deviance information criterion.

Finally, the BF_150 performed very well under the large DIF conditions (LG1 and LG2; non-invariance level) with almost 100% detection rates and small DIF conditions (SM1 and SM2; approximate-zero invariance) with detection rates 6% or below. Also, BF_150 detected the medium DIF (MD1 and MD2) fairly well (.98 to 1) when having a large number of groups. However, BF_150 has lower detection rates for the medium DIF magnitude conditions with the medium number of groups (GN=8). This is especially the case when the differences canceled each other out (MD1): the detection rates ranged between .30 and .33 in comparison to the range .72 and 1 for MD2.

In general, when BF, BIC and DIC were used to compare a model with .05 prior against a model with .01 prior, I observed no relation between the detection rates of Bayes factor and BIC although the former was calculated based on the latter. The performances of BF (BF_150) and DIC were comparable to each other with a slightly better performance of DIC for small DIF magnitude and a better performance of BF for medium DIF magnitude. When the cutoff points of BF were compared, the BF_150 worked better than BF_20 for small DIF magnitude. Even though BF_ 20 produced generally higher detection rates (between .49 and 1 compared to between .30 and .97 for BF_150) for medium DIF magnitude, the detection rates of BF_20 were high for small DIF (approximate invariance) especially with more groups (GN = 20).

Detection Rates of Approximate-Zero Scalar Invariance Testing across Five Prior Variances .001, .005, .01, .05, and .10 Using BIC, DIC and BF

For BIC and DIC, comparisons across five prior variances were conducted simultaneously and the model with the smallest value was selected as the best fitting model. The proportion of replications in which the approximate-zero scalar invariance was rejected (or the DIF was detected), that is, the model with a larger prior variance was selected, was computed for Bayes factor (BF).

BIC. Across all conditions, BIC kept selecting models with .10 prior, that is the biggest prior variance, over other prior models (see Table 17). This is not surprising because previous studies, (e.g., Kim et al., 2017), indicated that BIC tended to favor a model with a large prior variance. Table 17 presented the detection rats of BIC across the five prior variances.

Table 17
Selection Rates of Bayesian Approximate-Zero Scalar Invariance Tests across Five Prior Variances .001, .005, .01, .05, .10 Using BIC and DIC

Conditions			Model Comparisons Criteria					
GN	PCT	DIF-Size	BIC .10 Prior	DIC .001 Prior	DIC .005 Prior	DIC .01 Prior	DIC .05 Prior	DIC .10 Prior
20	50	LG1	1	0	0	0	.15	.85
		LG2	1	0	0	0	.15	.85
		MD1	1	0	.06	.81	.13	0
		MD2	1	0	0	.06	.94	0
		SM1	1	.98	.02	0	0	0
		SM2	1	1	0	0	0	0
	80	LG1	1	0	0	0	.42	.58
		LG2	1	0	0	0	.04	.96
		MD1	1	0	.64	.33	.03	0
		MD2	1	0	0	.69	.31	0
		SM1	1	1	0	0	0	0
		SM2	1	1	0	0	0	0
8	50	LG1	1	0	0	0	.57	.43
		LG2	1	0	0	0	.07	.93
		MD1	1	0	.02	.67	.28	0
		MD2	1	0	0	.23	.77	0
		SM1	1	.96	.01	.03	0	0
		SM2	1	.94	.02	.02	.02	0
	80	LG1	1	0	0	0	.60	.40
		LG2	1	0	0	0	.94	.06
		MD1	1	0	.21	.48	.31	0
		MD2	1	0	0	.41	.59	0
		SM1	1	.97	.02	0	.01	0
		SM2	1	.94	.02	.02	.02	0

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; BIC=Bayesian information criterion; DIC= Deviance information criterion.

DIC. The behavior of the DIC across the five prior variances was slightly different than its behavior when .05 prior to .01 prior variances models were compared. That is, when DIF was large, .05 or .10 was often selected; when DIF was small (approximate-zero invariance), .001

was by and large selected. Again, for medium DIF conditions (MD1 and MD2), the prior .01 was more often selected especially when item intercept differences canceled each other with a large number of groups (20).

Table 18

Detection Rates of Bayesian Approximate-Zero Scalar Invariance Tests with the Additional Pairs of Prior comparisons (.001 Prior vs. .05 Prior) and (.005 Prior vs. .05 Prior)

Conditions			DIC .001 Prior vs. .05 Prior	DIC .005 Prior vs. .05 Prior
GN	PCT	DIF-Size		
20	50	LG1	1	1
		LG2	1	1
		MD1	1	.45
		MD2	1	1
		SM1	0	0
		SM2	0	0
	80	LG1	1	1
		LG2	1	1
		MD1	1	.11
		MD2	1	.95
		SM1	0	0
		SM2	0	0
8	50	LG1	1	1
		LG2	1	1
		MD1	1	.62
		MD2	1	.98
		SM1	.03	.03
		SM2	.03	.04
	80	LG1	1	1
		LG2	1	1
		MD1	.98	.46
		MD2	1	.99
		SM1	.02	.01
		SM2	.03	.03

Note. GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction; DIC= Deviance information criterion.

Because the primary purpose of using five different prior variances was to figure out whether the conventional .01 and .05 comparison is reasonable or not, two additional pairs of prior comparisons were conducted: 1) a model with .001 prior against a model with .05, and 2) a model with .005 prior against a model with .05 prior. For .001 and .05 prior comparison, DIC

showed perfect performance: detecting large and medium DIF with almost 100% detection rates and supporting approximate scalar invariance when DIF was small. For .005 and .05 prior comparison, DIC showed good performance under the small and large DIF conditions with nearly 0% and 100 % detection rates, respectively. However, DIC has a moderate to low ability to detect both medium DIF conditions, (MD1 and MD2), especially when differences canceled each other out (MD1). The medium DIF detection rates were certainly low with .11 and .46 for large and medium unbalanced groups respectively. Tables 17 and 18 presented the DIC detection rates across the five prior variances and detection rates for .001 prior against .05 prior and .005 prior against .05 prior respectively.

Bayes factor (BF). Given that the BF is the ratio of two models' likelihood based on BIC, it is only able to compare two models at a time, (see Chapter 2 Equation (10) for BF calculation). In order to figure out whether the conventional .01 prior and .05 prior model comparison is reasonable or not, a total of five[2] pairs of model comparisons emerged: 1= prior .001 vs. prior .05, 2= prior .001 vs. prior .10, 3= prior .005 vs. prior .05, 4= prior .005 vs prior .10, 5= prior .01 vs. prior .10. The detection rates were expected to be very low and close to zero for small DIF-Size conditions, (i.e., SM1, SM2; approximate-zero invariance conditions) as favoring small prior models (.001, .005, .01), whereas they were expected to be high and close to 1 for moderate and large DIF-Size conditions, (i.e., MD1, MD2, LG1, LG2; non-invariance conditions) supporting large prior models (.05 and .10).

[2] Of note, not all the prior comparisons were applicable for all DIF conditions. For example, applying comparison (a model with prior .001 against a model with prior .005) to large and medium DIF conditions yielded invalid results because both prior variances were not acceptable with large DIF magnitude. Therefore, only valid comparisons based on the suitable prior variances were conducted and discussed.

Tables 19 and 20 presented the detection rates of BF for five new pairs of prior variances when BF_20 and BF_150 were used, respectively. The results of BF_3 are not presented or discussed because the behavior of the BF_3 across the five prior variances was consistent with its previous behavior when .05 prior to .01 prior variances models were compared.

Table 19

Detection Rates of BF_20 for BAMI Using Five Priors Models .001, .005, .01, .05, .10

Conditions			Bayes Factor Model Comparisons				
GN	PCT	DIF-Size	1 prior .001 vs. prior .05	2 prior .001 vs. prior .10	3 prior .005 vs. prior .05	4 prior .005 vs. prior .10	5 prior .01 vs. prior .10
20	50	LG1	1 [a]	1 [a]	1	1	1
		LG2	1 [a]	1 [a]	1	1	1
		MD1	1	1	1	1	1
		MD2	1	1	1	1	1
		SM1	1	1	1	1	.49
		SM2	1	1	1	1	.44
	80	LG1	1	1	1	1	1
		LG2	1 [a]	1 [a]	1	1	1
		MD1	1	1	1	1	1
		MD2	1	1	1	1	1
		SM1	1	1	1	1	.49
		SM2	1	1	1	1	.47
8	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	.89
		MD2	1	1	1	1	1
		SM1	1	1	.65	.64	.11
		SM2	1	1	.56	.56	.14
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	.76
		MD2	1	1	1	1	1
		SM1	1	1	.57	.59	.13
		SM2	1	1	.59	.59	.09

Note. BF_20= Bayes factor using BF ≥ 20 as cutoff value; GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction. [a]BF was undefined because the computed value was too big. I considered this as the selection of the model with a larger prior variance (DIF detected) because BF is greater than the cutoff.

Across the five pairs of model comparison, BF_20 performed poorly across all conditions: the detection rates under the approximate invariance (or small DIF) conditions were unacceptably high. BF_150 performed well under the large and small DIF conditions with a .01 prior against .10 prior models. However, low detection rates were observed for medium DIF conditions especially when item intercept differences canceled each other with a medium number

of groups (DIF-Size=MD1, GN=8). The behavior of the BF_150 using (.01 and .10) prior variances was similar to the behavior of (.05 and .01) prior variances pair. This finding suggests priors (.01 and .10) in additions to the (.01 and .05), that were both commonly adopted in previous simulation and applied studies, may provide reasonable results when BF with 150 as a cutoff value is used for model comparisons. Finally, the suitability of .001 and .01 priors for approximate-zero invariance was expected and supported by results of model fit assessment using PPP and 95% CI. Results showed that .001 and .01 priors perfectly fitted models with small DIF conditions.

Table 20

Detection Rates of BF_150 for BAMI Using Five Priors Models .001, .005, .01, .05, .10

Conditions			Bayes Factor Prior Pairs of Model Comparisons				
GN	PCT	DIF-Size	1 prior .001 vs. prior .05	2 prior .001 vs. prior .10	3 prior .005 vs. prior .05	4 prior .005 vs. prior .10	5 prior .01 vs. prior .10
20	50	LG1	1[a]	1[a]	1	1	1
		LG2	1[a]	1[a]	1	1	1
		MD1	1	1	1	1	1
		MD2	1	1	1	1	1
		SM1	1	1	.96	.96	.16
		SM2	1	1	.96	.96	.05
	80	LG1	1	1	1	1	1
		LG2	1[a]	1[a]	1	1	1
		MD1	1	1	1	1	.97
		MD2	1	1	1	1	1
		SM1	1	1	.92	.92	.09
		SM2	1	1	.93	.90	.05
8	50	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	1	1	.35
		MD2	1	1	1	1	.97
		SM1	1	1	.17	.18	.02
		SM2	.98	.98	.18	.19	.04
	80	LG1	1	1	1	1	1
		LG2	1	1	1	1	1
		MD1	1	1	.98	.98	.35
		MD2	1	1	1	1	.79
		SM1	.99	.99	.15	.20	.01
		SM2	.99	.99	.16	.16	.03

Note. BF_150= Bayes factor using BF ≥ 150 as cutoff value; GN= number of groups; PCT= percent of groups with non-invariant item intercept; DIF-Size= size and direction of item intercept differences; LG1= DIF magnitude of .6 in the non-invariant item intercept that cancel each other; LG2= DIF magnitude of .6 in the non-invariant item intercept with systematic direction ;MD1= DIF magnitude of .2 in the non-invariant item intercept that cancel each other; MD2=DIF magnitude of .2 in the non-invariant item intercept with systematic direction; SM1=DIF magnitude of.01in the non-invariant item intercept that cancel each other; SM2= DIF magnitude of.01in the non-invariant item intercept with systematic direction. [a]BF was undefined because the computed value was too big. I considered this as the selection of the model with a larger prior variance (DIF detected) because BF is greater than the cutoff.

Impacts of Simulation Design Factors

As shown in previous results, different sizes of impacts were observed across this simulation study. Given that no difference was found in terms of convergence rates in this simulation, it seemed that the four simulation factors: number of groups, percent of groups with non-invariant item intercepts, and direction and magnitude of item intercepts differences did not impact the convergence in this study. As expected, model fit became poor under large DIF conditions when either exact-zero or Bayesian approximate-zero scalar invariance model was fitted.

Impacts of number of groups (GN). Number of groups did not impact the results heavily because both group numbers (8 and 20) are common in cross-cultural research and sufficiently large. However, as oppose to GN=20, most of the variations in the results were associated with the medium number of groups (GN=8). The performance of BAMI model fit criteria were better when having medium group numbers (GN=8) rather than large (GN=20). Practically, the effect of the number of groups appeared more when it combined with unbalanced group (80% of groups with DIF items).

Impacts of percent of groups with non-invariant item intercept (PCT). PCT generally affected the detection rates, particularly when the DIF magnitude was medium for both exact and approximate MI tests. DIF detection rates were slightly higher for balanced conditions (50%) than unbalanced conditions (80%).

Impacts of magnitude and direction of item intercept differences (DIF-Size). This condition showed two levels of impacts: small impact was observed with the large and the small DIF conditions (LG1, LG2, SM1, and SM2) and, a more substantial impact was associated with medium DIF conations (MD1 and MD2). When DIF-Size were LG1 and LG2, or SM1, and

SM2, the effect seemed to be marginal because results were clear and straightforward, rejecting or supporting the Bayesian approximate-zero scalar MI. For the medium DIF magnitude conditions, (DIF-Size=MD1 and MD2), the impact of DIF directions became striking. The DIF detection rates were notably lower when item intercept differences canceled each other (MD1).

Cutoff Prior Precision Assessment

Results of this study are aligned with findings from previous Bayesian measurement invariance simulation and applied studies in classifying the five prior variances, (i.e., .001, .005, .01, .05, and .10), into two categories: approximate-zero invariance and substantial non-invariance level. A prior variance that equaled to .01 or less, (i.e., .001), was a representant of approximate-zero scalar invariance, whereas prior variance that equaled to .05 or greater (i.e., .10) was a representant of the substantial non-invariance level. Applying the previous priors classifications to the six DIF-Size conditions, the .001 and .01 prior variances are compatible with small DIF conditions (SM1 and SM2; approximate-zero invariance conditions) whereas .05 and .10 prior variances that are likely selected with large and medium DIF conditions (LG1, LG2, MD1, and MD2; non-invariance conditions).

Summary

This chapter described the results of the study. Twenty-six conditions under seven DIF scenarios were specified from which I obtained 100 datasets each, and then fitted into exact-zero scalar models using ML estimator, and approximate-zero invariance models using Bayes estimator. Both ML and Bayes estimators had the perfect convergence rates for all examined conditions. Regarding the model fit criteria, CFI and RMSEA, BIC, and AIC performed very well in correctly selecting models. For Bayesian models, both PPP and 95% CI performed very

well in detecting the suitable level of invariance that associated with the DIF magnitudes condition. Moreover, BF (with a cutoff 150) comparably performed to DIC in detecting the correct level of MI. They endorsed the approximate scalar invariance under the small DIF conditions whereas they detected the noninvariance well supporting metric invariance under the medium and large DIF conditions. In addition, the BAMI test at the scalar invariance level were found to have comparable results to the exact-zero scalar MI test with medium and large numbers of groups (8 and 20). Regarding the relationship between the simulated factors and the performance of the Bayesian approximate-zero scalar invariance tests, the percent of groups with non-invariant item intercept and the magnitude and direction of item intercept differences emerged as the most significant factors especially with a medium number of groups.

Chapter Five: Discussions

This chapter outlines the study main findings, discussion, followed by implications for the potential Bayesian measurement invariance researchers and methodologists. Then, limitations and future research are further discussed.

Main Findings[3]

The Performance of the Model Fit Criteria of the BAMI Testing in Detecting Non-Invariance Level

The first outcome variable to evaluate the performance was the convergence rates of the BAMI when applying to large and small balanced and unbalanced groups with three DIF magnitudes in two directions (total of 2*2*2*3=24 conditions). Each of these 24 conditions was fitted to the BAMI model with five prior levels: .001, .005, .01, .05, and .1, and categorized as models with small DIF (.001, .005, and .01 prior variances) and models with medium and large DIF (.05 and .10 prior variances). In the current study, Bayesian estimation reached convergence 100% across conditions.

The second outcome variable was the evaluation of the BAMI model fit criteria: PPP value larger than 0.05 and 95% CI including zero indicated a good model fit. PPP and 95% CI showed that models with .001 prior fitted data very well under both small DIF conditions but not well under other DIF-Sizes (i.e., LG1, LG2, MD1, MD2). Models with prior variances .005 and .01 fitted all small and medium DIF conditions except the large DIF-Size conditions (LG1 and

[3] I only summarized findings answered the two research questions. See Chapter 4 for through results

LG2). Finally, models with large prior variances (.05 and .10 prior) fitted all conditions regardless of the DIF magnitude or direction.

In additional to the PPP and 95% CI model fit evaluation criteria, three model comparisons criteria (mainly, Bayes factor (BF) at three cutoff points 3, 20, and 150, BIC, and DIC**)** were employed and the detection rates were reported. Overall, the BF using 150 and the DIC comparably performed in identifying the correct level of MI at scalar invariance testing with a better performance of DIC when having small DIF and a better performance of BF when having medium DIF. Finally, BIC failed to detect the correct MI level and tended to select the model with a larger prior variance (.05 or .10) even under small DIF conditions (approximate-zero invariance).

Both Bayes factor (cutoff point 150) and DIC supported the model with a small prior (.001, .01) under the approximate-zero scalar invariance conditions with .001 preferred by DIC only and .01 preferred by both DIC and BF_150). They also supported models with a larger prior variance (.05 and .10) under non-invariance conditions. However, under the medium non-invariance conditions in which the model with the .05 or .10 prior variance was expected to be selected, both DIC and BF_150 moderately performed especially under the large unbalanced groups when the differences canceled out each other.

In sum, four of the BAMI model fit criteria: PPP, 95% CI, BF_150, and DIC supported prior .01 to be a representant of approximate-zero invariance. Moreover, three of the BAMI model fit criteria: PPP, 95% CI, and DIC supported prior .001 to be a representant of approximate-zero invariance. However, BIC failed to support both priors or any small prior and, therefore, recommended to be excluded when using BAMI approach for testing MI.

The Impacts of the Design Factors on the Simulation Outcomes of Testing and Estimating the Approximate Measurement Invariance

The simulation factors: group numbers, percent of groups with non-invariant item intercept, and direction and magnitude of difference in item intercepts did not impact the convergence in this study. In general, BAMI showed different results and the GN and the DIF-Size emerged as the most significant factors. However, with a large group number, the DIF-Size was the only factor that moderately influenced the results. The effect of medium number of groups appeared stronger when it is combined with 80% of groups with DIF items. BAMI test showed more consistent results across simulation factors under large and small DIF conditions. However, the medium DIF magnitude was less well detected especially when item intercept differences cancelled each other in the unbalanced groups.

Finally, based on the variations in results that associated with the employed prior values, a few recommendations of cutoff prior precision value emerged. First, with a large group number, using .01 prior variance to provide wiggle room that could robustly handle a small discrepancy in item intercepts across groups worked very well with balanced and unbalanced groups using both DIC and BF_150 as model comparisons criteria. Second, prior .001 worked very well to reflect Bayesian approximate-zero invariance at the scalar level with small or large balanced and unbalanced groups when DIC was the only model comparison criterion. A prior variance larger than .01 was proved to be substantial non-invariance and of course not endorsed for fitting an approximate-zero scalar MI. These recommendations were based on the study setting such as large sample size per group, around 500, and a decent number of groups, preferably balanced.

Discussions

Muthén and Asparouhov (2012a, 2013) introduced the approximate level as new concept of measurement invariance that was based in Bayesian estimates rather than in the traditional frequentist approach. The Bayesian approximate MI testing was seen as an alternative approach to override the sensitivity of the strict exact-zero invariance assumption for trivial non-invariance. Applied and simulation studies have used the BAMI across different numbers of groups. Although research showed that the BAMI had promising results, a systematic review for applied and simulation research using BAMI showed that the BAMI method has not been well explored across several research settings (see Chapter 2). Therefore, this study was driven by the need to delineate some guidelines of the BAMI approach including a prior size, the acceptable bias size, total number of groups, and percentage of groups that have DIF items. Additionally, this study aimed to help in setting up the "golden" rules for evaluating model fits through examining the behavior of five model fit criteria, two criteria for evaluating the model fit and three criteria for model comparisons. Based on the analytic comparisons and simulation results in this study, four major findings emerged.

First, the BAMI approach is appropriate, as MI testing model, to provide a valid MI testing results if a suitable pair of prior variances that are combined with the appropriate model comparison criterion are used, which requires a good level of knowledge about priors and criterion behavior, pros, and cons across different research settings. In this study, the BF, BIC, and DIC were used to evaluate the BAMI approach using (.01 and .05) pair of priors. A prior variance that equaled to .01 was a representant of approximate-zero scalar invariance whereas prior variance that equaled to .05 was a representant of the substantial non-invariance level. Exploring the Bayes factor (BF) provides an insight to the BAMI literature (BF $\geq$ 3, BF $\geq$ 20,

and BF ≥ 150). Two of the assigned BF cutoff points, 20 and 150 showed a moderate level of agreements in their results with a better performance of BF_150 over BF_ 20. Meanwhile, the results of BF_150 are endorsed by the DIC results. Both BF_150 and DIC were able to detect the correct BAMI models to a close degree. Unfortunately, the BIC failed to detect the correct BAMI model based on the simulation conditions and the appropriate prior size. In every model comparison, the BIC tended to select the large prior, which was not surprising because it aligned with previous research, (e.g. Kim et al., 2017). Interestingly, no relation was observed between the performances of Bayes factor and BIC although the difference between BIC values transforms to an approximation of the Bayes factor. However, as seen in Equation 10 (see Chapter 2), the BIC absolute values are irrelevant—only the differences in BICs carry evidential weight. The outperformance of the BF over BIC also was observed in previous research (see Wagenmakers, 2007). With this in mind, using BF with 150 cutoff point or/and DIC is recommended with (.01 and .05) pair of prior variances to produce trustworthy BAMI model comparisons results.

Previous simulation and applied studies used prior values of (.001), (.005), and (.01) as approximate-zero DIF (small) whereas (.05) and (.10) were considered as non-invariance level (medium and large DIF). The reported pair of prior variances that represented the two cutoff priors for small (approximate-zero) and large DIF (non-invariance) were different per study. Therefore, one of the main purposes of using five different prior variances in this study was to figure out whether the conventional pair of priors (.01 and .05) comparison is reasonable or not. Additional pairs of prior comparisons were conducted, and three pair of prior variances were suggested, (i.e., .001 and .05, .01 and .05, and .01 and .10). These results are supported by previous research. For example, across the 10 BAMI applied research, (see Chapter 2 Table 4 for

details), four studies used (.01 and .05), two studies used (.005 and .05), and two studies used (.01 and .10). Regarding the three BAMI simulation studies, (i.e., Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013), the pair priors (.001 and .05) were only endorsed by Kim et al. (2017). Also, Muthén and Asparouhov (2013) was the only study that supported pair priors (.01 and .10). However, both Muthén and Asparouhov (2013) and van de Schoot et al. (2013) suggested pair prior variances (.01 and .05). However, the suggested three pairs of priors, (i.e., .001 and .05, .01 and .05, and .01 and .10), by this study were conditional to a specific model comparison criterion: for example, using DIC and BF_150 with (.01 and .05), using only DIC with (.001 and .5), and using only BF_150 with (.01 and .10). To put it another way, the three proposed pairs of prior variances by this study seem to be practical if the suitable model comparison criterion was used. Kim et al (2017) supported the DIC with (.001 and .05), but they considered a smaller size of noninvariance as approximate (.009 prior variance). However, the usage of BF with 150 as a cutoff for BAMI is newly introduced by this study and more research are needed for generalization of the results. Finally, of note is that the prior variance .01 corresponded to the generated DIF magnitudes for small noninvariance in this study. Thus, if a researcher pre-determines the approximate MI or the tolerable size of noninvariance (e.g., .01 variance or smaller) and use this as a cutoff in BAMI testing (e.g., .01 vs. .05), it is generally expected that a correct level of MI is detected if appropriate model comparison indices such as DIC and BF_150 are used.

Second, because the BAMI is a Bayesian approach, researchers ought to make several important decisions in advance based on their own experiences, experts' advices, or previous research in the field. These decisions would positively (or negatively) affect the quality of the results (see Chapter 2). For this study, the prior size and the size of the acceptable DIF, as

approximate-zero intercepts differences across groups, were the most important decisions. Based in the BAMI literature, five levels of prior precisions, (.001, .005, .01, .05, and .10), and three DIF magnitudes were used as small (.01), medium (.2) and large (.6). The PPP> .05 and the 95% CI encompass zero were used as indicative of good model fit. The .001 prior was sensitive to medium and large DIF magnitudes; the .005 and .01 priors were sensitive to large DIF; the .05 and .10 priors were insensitive to all DIF magnitudes in this study. These results were supported and found in previous BAMI simulation and applied studies because a prior variance that equaled to .01 or less, (i.e., .001, .005) was a representant of approximate-zero scalar MI whereas prior variance that equaled to .05 or greater (i.e., .10) was a representant of the substantial non-invariance level (see Chapter 2). The BAMI model fits results showed that the PPP and the 95% CI were reasonable criteria for model evaluation except the two following situations/exceptions.

The first and most frequent situation was when the DIF magnitude is medium, (where item intercept differences size = .2). The medium DIF magnitude showed inconsistent reactions toward the BAMI model**,** especially when differences canceled each other in unbalanced group number. The medium size of DIF was generated to fit the large prior size, (.5 and .10 priors). Yet, it produced a paradox in the results because it often selected .005 or .01 prior in particular. Given that the prior variance of .01 (.1 SD) allowed wiggle room between -.2 and .2 (±2 SD) for intercept differences between groups, a reason for low detection rates with this DIF magnitude is the subjectivity in the decision that made about the size of approximate invariance and the generated size of medium DIF (.2).

The second exception was when the assigned prior is larger than the DIF magnitude, (e.g., when .05 prior applied to small DIF conditions). An extreme PPP value (PPP< .05) was expected when the prior variance is small relative to the magnitude of non-invariance, so the PPP

will not belong to the distribution of the correctly specified model and it is in the tail of the distribution (Muthén & Asparouhov, 2013). However, the extreme PPP value was not observed because the prior variance (or the allowed wiggle room for the DIF in the intercept estimates) was larger than the magnitude of non-invariance, therefore, the PPP and the 95% CI failed to detect the misfit even though the assigned prior was not suitable to the small DIF magnitude. While it is not the focus of this study**,** as a side note, increasing the prior variance will not affect the model fit results, (PPP and 95% CI), but only will affect parameter estimates and sizes of standard errors.

Third, as observed in previous simulation research, (e.g., Kim et al., 2017; Muthén & Asparouhov, 2013; van de Schoot et al., 2013), related factors in the research setting such as group numbers, percent of groups with non-invariant items intercepts, and DIF items magnitude and directions may affect the performance of the BAMI. In general, DIF directions in item intercepts (i.e., cancel each other out and systematic) impacted the BAMI results. In one hand, some cases with systematic DIF directions performed poorly when it combined with unbalanced DIF number of groups (16 out of 20 groups and 6 out of 8 groups). On the other hand, the cases with DIF differences canceled each other notably performed poorly. Moreover, the medium number of groups (GN=8) showed a variation across the results. The effect of the medium number of groups factor increased when it combined with the unbalanced groups (6 groups with DIF items out of 8). One explanation of these variations is that with a smaller sample, the prior has a stronger effect (Muthén, 2010; Yuan & MacKinnon, 2009).

Fourth, to determine whether or not the Bayesian approximate scalar invariance holds, the fit of a model with a predetermined prior is compared against models with several priors variances. In this study, when applying the BAMI scalar models, two prior variances showed

promising results for small DIF, (.001 and .01 prior variances; approximate-zero). The prior variance of .001 (.032 SD) allowed wiggle room between −.064 and .064 for intercept differences between groups whereas the prior variance of .01 (.1 SD) allowed wiggle room between -.2 and .2 for intercept differences between groups. However, the choice for a prior variance is crucial to the BAMI because the BAMI is only able to correctly identify the approximate-zero level if the selected prior is suitable for the size of the parameter's differences across groups. Consequently, finding of this study suggests that the use of default or diffuse (noninformative) prior is not recommenced with BAMI models. "The use of informative priors clearly needs to be approached with caution. An investigator must not choose a certain prior because it makes it more likely to find an intervention effect" (Muthén, 2010, p.12). Researchers require to apply their knowledge on the distribution of item parameter non-invariance and predicting the suitable size of prior. Of note, researchers who have insufficient priors information, they may obtain informative priors by seeking advices from the experienced researchers or from previous studies in a field.

Definitely, the Bayesian approximate measurement invariance is a good tool for MI testing, however, the exact-zero level of invariance is not attainable by using BAMI. The BAMI can be used by the only researchers who opt for the approximate-zero invariance level. Results of this study indorsed that the optimal usage for the BAMI is when the number of groups is large with many small parameters differences which was recommended by previous BAMI studies such as van de Schoot et al. (2013) and Kim et al. (2017).

Implications

Potential researchers, who are interested in testing for measurement invariance, may opt to select a method from different MI approaches. However, the study setting, the sufficient

information about scales, and the intended usage/ purpose of conducting the model comparison across groups would determine the suitable approach. When researchers intend to establish approximate-zero invariance, the Bayesian MI method is recommended. There is no confirmed cutoff or rule of thumb about when to use a specific MI approach and why, but results of this study can inform decisions regarding BAMI across many groups and provided some implications and recommendations for the appropriate use.

First, this study focused on the performance of the BAMI under two model fit criteria and three model comparisons indices. It provided the evidence that the BAMI method is robust and able to correctly detect the invariance level under 24 conditions when a suitable prior variance and fit criterion were used; even though the study only considered the non-invariance in the item intercept. Researchers conducting a BAMI analysis hence might consider applying the BAMI method in testing the approximate-zero invariance in both loadings and intercepts parameters to fulfill goals for estimation and inference that simply cannot be accomplished by existing frequentist approaches. Although there is no clear adherence to adequate rules for the cutoff point of the Bayes factor, the explorations of the trio cutoff points, 3, 20, and 150, of the performance of the Bayes factor enhanced the quality of the BAMI results and supported the results of other BAMI indices, i.e., DIC. Hence, future researchers are able to use BF with a cutoff of 150, with a solid foundation of the efficiency of BF with the BAMI approach along with DIC when their study conditions are similar to those in this simulation. Future research is also called for to second the current results of the BF using the three cutoff points by applying them to a new research scenario.

When BF is used, one must be aware of the sensitivity of BF to model with improper prior, for example, fitting a model with a very small prior variance to large DIF (Hooten &

Hobbs 2015; Spiegelhalter & Smith, 1982). “It is well known that the resulting Bayes factor involves an arbitrary, unspecified constant, and is thus not well defined” (Spiegelhalter & Smith, 1982, p. 377). When an unrealistically small prior variance (i.e., .001) was applied to large DIF, BF was undefined because this model fit was too poor to compute BF, and I considered this case as a detection of DIF. Applied researchers should be cognizant of undefined BF when the fit of one model is expected to be too poor (e.g., a very small prior variance when DIF is large). Calculating BF using method that is not involving exponential function might be considered. One way is referring to Kass and Raftery (1995) concordance table that equalized Bayes factor to the difference between BIC of two competing models. For example, a BF cutoff point of 150 is equal to a BIC difference of 10, allowing for a strong posterior probability that the competing model is the preferred model.

Second, this study aimed to quantify the approximate-zero deviating between parameters (DIF magnitude) that is combined with the approximate variance (prior). Therefore, this study compared the performance of the BAMI under five prior levels in order to define the “approximate” MI level that associated with the “small” DIF magnitude. Emphasizing the prior size that was able to accommodate the wiggle room for parameter differences and produced a good BAMI model fit, the BAMI approach with priors between .001 and .01 appears reasonable if DIC is used for model selection. A smaller prior variance such as .001 with DIC could detect the medium size noninvariance while it supported approximate scalar invariance when the DIF magnitude was small. However, small priors such as .001 and .005 are not recommended with BF because approximate invariance was almost always rejected even when approximate invariance was generated. For BF, the pair (.01 and .10) can be considered instead if a researcher concerns a medium size DIF. Overall, the original prior pair (.01 and .05) is recommended for

both DIC and BF_150 especially if the situations where medium size differences cancelled out each other are not of concern in a study. This might help future researchers to find the definition of approximate-zero invariance, based on the size of prior and DIF, which was vague across the BAMI applied studies (see Chapter 2).

Furthermore, not only Bayesian researchers may benefit from quantifying the approximate or small DIF but also researchers who opt to use the frequentist approaches (exact-zero MI testing) do. As seen in the BAMI applied literature, some researchers criticized that the exact-zero scalar invariance test (that is, allowing for zero differences in item loadings and intercepts across groups) could be too sensitive to trivial non-invariance. However, in this study, when approximate-zero scalar invariance was generated in the population, the detection rates that falsely detected DIF and rejected exact scalar invariance using goodness-of-fit statistics, namely, LRT, ΔCFI, ΔRMSEA, BIC, and AIC, were found to be less than 5% except LRT. These indices except BIC were more sensitive to the medium size of DIF in this study relative to the fit indices of BAMI tests. This seems to be promising for researchers who opt to use the exact MI testing if model comparison methods such as ΔCFI are used.

Limitations and Directions for Future Study

Given the research design, like other simulation studies, this study has several limitations in its scope. First of all, this study focused only on applying the BAMI approach to the item intercepts differences, and therefore, tested only for scalar MI. Although the scalar invariance level (as an advanced level) has been frequently used in previous BAMI studies and has been known as a challenging level in cross-cultural research, there are a wide variety of models in practice varying or changing the differences in both loadings and intercepts parameters. As shown in Chapter 2, testing the Bayesian approximate-zero invariance at the metric level could

be done before establishing the approximate-zero scalar invariance. Models with a mix of parameters differences, in items intercepts and loadings, could be examined for further investigation.

Additionally, the CFA model used in this study had a one latent variable with six indicators. In practice, more complex models are present and thus the approximate prior variance investigated in this study can be limited in reality. The Bayesian has an advantage for estimating more complex models, and hence, in order to examine whether the results of the current study can be generalized to a variety of models, various SEM models could be examined for further investigation.

There are also important simulation design factors not manipulated in the current study which deserve some attention, particularly in a Bayesian analysis. This study included only fixed sample size across all groups, that is 500. In reality, there are many unequal sample sizes across groups. A good example of an unequal sample size situation is in the international large-scale assessments, the Trend of International Mathematics and Science Study (TIMSS) for example, where the participated countries differed in their sizes and therefore so did the number of students. In TIMSS 2015, 8^{th} grade mathematics section, the total number of participants from Saudi Arabia were 3759, Oman students were 8883, whereas the United Arab of Emirates participants were more than 18012 students. Unequal sample sizes across groups may have an impact on statistical conclusions and inferences on the BAMI model comparisons across groups. Therefore, additional factors for different sample size could be investigated for further study. By doing so, the advantages and disadvantages of using small prior value such as .001 or .01 on the item intercept differences could also be better examined.

Moreover, not only could unequaled sample size be manipulated, but also examining small sample size could be manipulated. Although the intended research for this study was the cross-cultural research, which usually have large sample size, the BAMI approach may apply to any number of groups or sizes. More specifically, research showed that Bayesian worked very well with small sample size, and with a smaller sample, the prior has a stronger effect. Therefore, future researchers are encouraged to include small size of groups. By doing so, the cutoff prior value to define as approximate that is recommended by this study could also be reevaluated.

Nonetheless, the results of this study provide valuable information about to what extent the BAMI approach was robust to different research settings based on different model fit criteria and how these setting impacted the BAMI results. The author hopes the study will allay concerns about the use of informative priors as "approximate" in cross-cultural research and in doing so encourage a high level of reporting transparency because priors need to draw on information from other similar studies.

Finally, the results of the current study guide the potential study for thoroughly examining the sitting of the BAMI model evaluation rules and fit criteria along with expanded simulation conditions. Also, exploring the quality of the Bayes factor performance with its three cutoff points enhanced and indorsed the quality of the performances of other model fit criteria, (e.g., DIC), and therefore, future researchers would be more affirmative about scales invariance decisions and interpretations across their groups. Moreover, even though the performance of the frequentist approach and the Bayesian approach in MI for group comparison was discussed in previous research, the use of the Bayes factor as a new model fit criterion in BAMI and results of the current study along with expanded simulation conditions would help future researchers to be more knowledgeable about the strengths and weakness of the two approaches. Because the

BAMI is mainly targeting research of cross-national studies with a large group number, there could be other BAMI research and issues, for example the impact of data missingness. An interesting line of research is to study the behavior of the BAMI approach under different patterns of missingness. This might be an interesting study because the cross-cultural or cross-national studies data usually have high rates of missingness because they are not data for high-stake assessment.

References

Articles included in the review of applied research (Chapter 2) are indicated with an asterisk (*)

American Educational Research Association [AERA], American Psychology Association [APA], & National Council on Measurement in Education [NCME]. (2014). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology, 66*(1), 1-7. doi: 10.1111/bmsp.12004

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis using M*plus*: Technical implementation. Retrieved from www.statmodel.com/download/Bayes3.pdf

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 495-508. doi:10.1080/10705511.2014.919210

Bandalos, D., L., & Gagne, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.). *Handbook of structure equation modeling* (pp. 92-108). London; The Guilford Press.

Bayes, T., Price, R., & Canton, J. (1763). An essay towards solving a problem in the doctrine of chances. In C. Davis (Ed.), *Philosophical Transactions* (pp. 370-418). London: The Royal Society of London.

Beierlein, C., Davidov, E., Schmidt, P., Schwartz, S. H., & Rammstedt, B. (2012). Testing the discriminant validity of Schwartz' Portrait Value Questionnaire items: A replication and extension of Knoppen and Saris 2009. *Survey Research Methods*, *6*, 25–36. doi:10.18148/srm/2012.v6i1.5092#sthash.P8YCGK8T.dpuf

*Braeken, J., & Blömeke, S. (2016). Comparing future teachers' beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning. *Assessment & Evaluation in Higher Education*, *41*(5), 733-749. doi:10.1080/02602938.2016.1161005

Braeken, J., Mulder, J., & Wood, S. (2015). Relative effects at work: Bayes factors for order hypotheses. *Journal of Management*, *41* (2), 544-573. doi: 10.1177/0149206314525206

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Publications.

*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), 1-10. doi:10.3389/fpsyg.2014.00984/full

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

Byrne, B. M., & van de Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10* (2), 107-132. doi: 10.1080/15305051003637306

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464-504.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*, 233-255. doi:10.1207/S15328007SEM0902_5

Christensen, R., Johnson, W., Branscum, A., Hanson, T. E. (2011). *Bayesian ideas and data analysis: An introduction for scientists and statisticians.* Boca Raton, FL: CRC Press.

*Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact versus an approximate (Bayesian) measurement invariance test: A cross-country illustration with a new scale to measure 19 human values. *Frontiers in Psychology*, *5* (982), 1-10. doi:10.3389/fpsyg.2014.00982/full

Cronbach, L.J., (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16* (3), 297-334. doi:10.1007/BF02310555

*Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M. (2015). The comparability of measurements of attitudes toward immigration in the European social survey exact versus approximate measurement equivalence. *Public Opinion Quarterly*, *79* (S1), 244-266. doi:10.1093/poq/nfv008

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*, 55-75.

*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*. doi:10.3389/fpsyg.2015.01963

Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-checklist. *Psychological Method, 22* (2), 240-261. doi:10.1037/met0000065

Desa, D. (2014). *Evaluating measurement invariance of TALIS 2013 complex scales: A comparison between continuous and categorical multiple-group confirmatory factor analyses* (EDU/ WKP [2014] 2). Paris, France: OECD Publishing.

Elsworth, G. R., Beauchamp, A., & Osborne, R. H. (2016). Measuring health literacy in community agencies: A Bayesian study of the factor structure and measurement invariance of the Health Literacy Questionnaire (HLQ). *BMC Health Services Research*, *16*, 508. doi:10.1186/s12913-016-1754-2

French, B. F., & Finch, H. W. (2008). Multigroup confirmatory factor analysis: Locating the invariant referent sets. *Structural Equation Modeling: A Multidisciplinary Journal*,*15*(1), 96–113. doi:10.1080/10705510701758349

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, (4), 457-472.

Gelman, A., Carlin, J. B., Stern, H.S. & Dunson, D.B. (2014). *Bayesian data analysis* (3rd ed.). Boca Raton: FL: CRC Press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, (4), 733-760.

Gill, P. S., & Swartz, T. B. (2004). Bayesian analysis of directed graphs data with applications to social networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *53*(2), 249-260.

*Gucciardi, D. F., Zhang, C. Q., Ponnusamy, V., Si, G., & Stenling, A. (2016). Cross-cultural invariance of the mental toughness inventory among Australian, Chinese, and Malaysian athletes: A Bayesian estimation approach. *Journal of Sport and Exercise Psychology*, *38*(2), 187-202.

*He, J., & Kubacka, K. (2015). Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013. (OECD Education Working Papers, No. 124). Paris, France: OECD. doi:10.1787/5jrp6fwtmhf2-en

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, *2*(2), 1-19. doi: doi.org/10.9707/2307-0919.1111

Hoff, P. D. (2009), *A First course in Bayesian statistical methods*. Springer Science & Business Media.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367.

Hooten, M.B., & Hobbs, N.T. (2015). A guide to Bayesian model selection for ecologists. *Ecological Monographs*, 85(1), 3–28.

Hue, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55. doi:10.1080/10705519909540118

Jackman, S. (2000). Estimation and inference via Bayesian simulation: an introduction to Markov Chain Monte Carlo. *American Journal of Political Science*, *44* (2), 375-404. doi:10.2307/2669318

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

Kaplan, D. & Depaoli, S, (2012). Bayesian structure equation modeling. In Hoyle, R.H., (Ed.). *Handbook of structure equation modeling* (pp. 650-673). Guilford Publications.

Kass, R. A, & Raftery, A. E. (1995). Bayes factors. *Journal of The American Statistical Association*, (430), 773. doi:10.2307/2291091

Kim, E., S., Cao, C., Wang, Y., & Nguyen, D., T. (2017). Measurement invariance testing with many groups: A comparison of five approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 524-544. doi:10.1080/10705511.2017.1304822

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

Lee, P., (2012). *Bayesian statistics: An introduction* (4th ed.). West Sussex: UK, John Wiley & Sons Ltd.

Lunn, D.J., Thomas, A., Best, N., & Spiegelhalter, D. (2000) WinBUGS: A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician), 49* (3), 293-337.

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution, 3*(1), 112-115. doi: 10.1111/j.2041-210X.2011.00131.x

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling*, *13*(1), 59-72. doi:10.1207/s15328007sem1301_3

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *11* (4), 514-534.

MacEachern, S., & Berliner, L. (1994). Subsampling the Gibbs sampler. *The American Statistician, 48*(3), 188-190. doi:10.2307/2684714

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum Associates.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(4), 611–635. doi:10.1080/10705510701575461

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

Milfont, T. L. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111-121.

Mulder, J., & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments." *Journal of Mathematical Psychology*, 72, 1–5, doi: 10.1016/j.jmp.2016.01.002

Muthén, B. O. (2010). Bayesian analysis in M*plus*: A brief introduction. *Unpublished manuscript.* Retrieved from: https://www.statmodel.com/download/IntroBayesVersion%203.pdf

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

*Muthén, B., & Asparouhov, T. (2013). BSEM measurement invariance analysis. M*plus Web Notes*, *17*, 1-48. Retrieved from https://www.statmodel.com/examples/webnotes/webnote17.pdf

Muthén, B. O. & Asparouhov, T. (2012b). New Developments in Mplus Version 7: Part 1 [PowerPoint slides]. Retrieved from https://www.statmodel.com/download/handouts/V7Part1.pdf

Muthén, B. O. (2002). Using M*plus* Monte Carlo simulations in practice: A note on assessing estimation quality and power in latent variable models. *Mplus Web Notes, 1*.

Muthén, B. O., & Muthén, L. K. (2017). M*plus* 8.0 [Computer software]. Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Nagengast, B., & Marsh, H. W. (2013). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup SEMs across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment*, (pp. 317-344). Boca Raton, FL: Chapman & Hall, CRC Press.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8*(2), 287-312.

Raftery, A.E., & Lewis, S.M. (1996). Implementing MCMC. In W., R. Gilks, D.J. Spiegelhalter, & S. Richardson (Eds.), *Markov chain monte carlo in practice* (pp.115-130). London: Chapman & Hall.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57. doi:10.1177/0013164413498257

Sass., D. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29*(4), 347– 363.

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

Spiegelhalter, D. J. & Smith, A. F. M. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society Series B (Methodological), 44* (3), 377–387

van de Schoot, R., Schmidt, P., de Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology, 1064* (6). doi:10.3389/fpsyg.2015.01064/full

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist. 16* (2), 75-84.

*van de Schoot, R., Kluytmans, A., Tummers, L. G., 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, 770* (4), 1-15. doi: doi.org.ezproxy.lib.usf.edu/10.3389/fpsyg.2013.00770

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217. doi: dx.doi.org.ezproxy.lib.usf.edu/10.1037/met0000100

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. doi: 10.1177/109442810031002.

Verhagen,J., Levy, R., Millsap, R., & Fox, J. P. (2016). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*. 72, 171-182. doi: 10.1016/j.jmp.2015.06.005

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779-804. doi:10.3758/BF03194105

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. *Psychological Methods*, *14*(4), 301–322. doi:10.1037/a0016972

*Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E. (2015). The comparability of the universalism value over time and across countries in the European Social Survey: exact vs. approximate measurement invariance. *Frontiers in Psychology*, *6*, 733. doi: 10.3389/fpsyg.2015.00733

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, *41*(2), 390-420. doi:10.1177/0149206313501200

Appendix A. Bayesian Approximate Measurement Invariance (BAMI) Coding Protocol

Title:		*Journal:*			
Year:		*Volume (issue), pages:*			
Author (s):					
Item	**Yes**	**No**	**Page**	**Comment**	
1. Is there an appendix and/or endnote provided with BAMI details?					
Study Characteristics					
2. Purpose *2.a. Test for measurement invariance* *2.b. Compare MI and AMI* *2.c. Both* *2.d. Other: List……*					
3.Design: 3.a Simulation 3.b. Applied 4. BAMI Framework: 4.a. Muthén & Asparouhov, 2013 4.b. Schoot et al., 2013 4.c. Other: List…… 5. Provide rationale of: 5.a. Bayesian: List…….. 5.b. Approximate MI: List……... 6. Address MI as a main research question or as one of the main research Questions.					
Scale *7. Scale Name* *8. Developer and date* *9. Factor Numbers* *10. Items Numbers* *11. Response Scale* *12. Discussed Scale Translation* *13. Reported scale factors dimensions* *14. Provide a copy of scale items*					
Data 15. Type 15.a. Continuous 15.b. Ordinary 15.c. Dichotomous 15.d. Treated as continuous: Explain……..……... 15.e. No data information 16.Missing Data discussed 17.Normality assumption discussed					

18. Summary of descriptive statistics provided
19.Data collection
    19.a. Cross Sectional
    19.b. Secondary Data
    19.c Demo Purpose
20. Sampling method: List……
21.Sample Size
    21.a. Total
    21.b. Sample size per group ranged …….to……
    21.b. Sample size justified
22. Number of Groups……..
23.Administration Formats
    23.a. Online
    23.b. Written
    23.c. Both
    23.d. No information
24. Participants information
    24.a. Gender
        24.a.1 Total number per type
        24.a.2 By percent
    24.b Age
        24.b.1 by interval
        24.b.2 by percent
    24.c Nationality
    24.d Education
    24.e. Other: List……..…..
    24.f. No information

Software Program

25. Bayesian Software program
    25.a. Mplus
    25.b. OpenBUGS
    25.c. R
    25.d. Other: List……
    25.e. Version ( )

26.Code
  26.a As part of study
      26.a.1 Complete
      26.a.2 Partial
  26.b Code as a supplementary material
      26.b.1 Complete
      26.b.2 Partial
27. Type
28. Model (allfree)
29. Chains: Number…..…….
30. Thin: Every ……
31. Fbiter: Number…..…….
32. Processor: Number…..…….
33. Biterations
    33.a. Minimum …..…….
    33.b. Maximum …..…….
34. Burn-in: Number…..…….
35. Bconvergence …….
36. Bseed ……..

37.MGCFA (groups/time)
37.a. Cross-Cultural
37.b. Cross-Country
37.c. Cross Gender/ Cross-Group: List……..
37.d Time point: List ……..

Model Specification & Estimation

38. Model Identification
    38.a. Marker-variable(s)
    38.b. Standardized factor(s)
39. Test MI before BAMI
    39.a. Full MI invariance level hold…….
    39.b. Partial MI invariance level……
40. Number of bias item before BAMI
41. Number of MI models tested
42. BAMI procedure
    42.a Sequential MI testing
    42.b Testing metric and scalar simultaneously
    42.c Testing only two levels: configural and then Scalar
    42.d. Other: List…….
43. Algorithm
    43.a MCMC Gibbs sampler
    43.b. Other: List…….
44. Prior
    44.a. Specified
        44.a.1 One prior for all items
        44.a.2 Other: List…….
    44.b. Justified
    44.c. Source provided: List…….
45. Prior type
    45.a. Informative
    45.b. Noninformative
46. Number of priors that used
47. Prior distribution for factor means and variances
48. Prior distribution difference in loadings
49. Prior distribution difference in intercepts
50. Residuals correlated/ uncorrelated
51. Prior for residual covariances
52. Prior distribution difference in residuals variance

Convergence & Model Evaluation

53. Convergence
    53.a Visual inspection
        53.a.1 trace plot
        53.a.2 autocorrelation
    53.b Statistically
    53.c. Cut-off value for convergence: List …….
    53.d. Non-convergence discussed
54. Bayesian fit indices
    54.a. PPP
        54.a.1 Cut off: List …….
    54.b. 95% credibility interval of χ^2 includes zero
    54.c. BIC
    54.d. DIC
    54.e. Deviance
    54.f. Other: List ……..

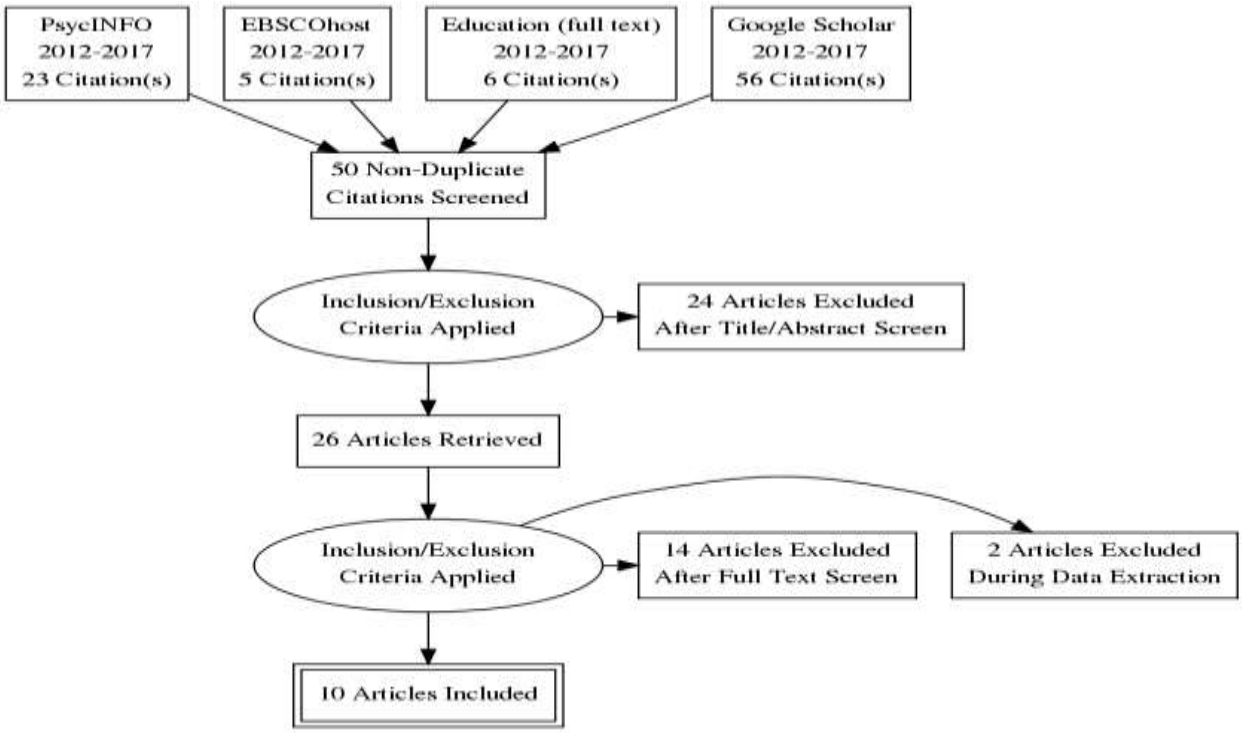
Results	*55. Level of MI hold (after BAMI)* *55.a. Full MI* *55.b. Partial MI*
	56. Number of bias items after BAMI
	57. Presentation
	57.I. Diagram 58.a Scale diagram 58.b. Prior Distribution 58.c Bayesian plots (trace plot/ autocorrelation) 58.c. Approximate MI 58.e. N/A 58.f. Other: List…….
	57.II. Table 58.a. Correlations matrix 58.b. Summary of Descriptive Statistics 58.c. BAMI fit indices 58.d. BAMI parameters: 58.d.1 List ……. 58.d.2 STDY 58.f. Partial BAMI 58.g. Bias Items 58.h. Groups associated with bias items 58.i. N/A 58.j. Other: List…….

Appendix B. Summary Table for Information of the Reviewed Articles

Assigned Number	Article title	Authors	Year	Journal
1	Psychometric evaluation of the overexcitability questionnaire-two applying Bayesian structural equation modeling (BSEM) and multiple-group BSEM-based alignment with approximate measurement invariance	De Bondt, N., & Van Petegem, P.	2015	Frontiers in Psychology
2	The comparability of measurements of attitudes toward immigration in the European social survey exact versus approximate measurement equivalence	Davidov, E., Cieciuch, J., Meuleman, B., Schmidt, P., Algesheimer, R., & Hausherr, M.	2015	Public Opinion Quarterly
3	Measuring hedonia and eudaimonia as motives for activities: cross-national investigation through traditional and Bayesian structural equation modeling	Bujacz, A., Vittersø, J., Huta, V., & Kaczmarek, L. D.	2014	Frontiers in Psychology
4	Comparing results of an exact vs. An approximate Bayesian measurement invariance test: a cross-country illustration with a scale to measure 19 human values	Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H.	2014	Frontiers in psychology
5	Comparing future teachers' beliefs across countries: approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item functioning.	Braeken, J., & Blömeke, S.	2016	Assessment & Evaluation in Higher Education
6	The comparability of the universalism value over time and across countries in the European social survey: exact vs. Approximate measurement invariance	Zercher, F., Schmidt, P., Cieciuch, J., & Davidov, E.	2015	Frontiers in Psychology
7	Cross-cultural invariance of the mental toughness inventory among Australian, Chinese, and Malaysian athletes: a Bayesian estimation approach	Gucciardi, D. F., Zhang, C. Q., Ponnusamy, V., Si, G., & Stenling, A.	2016	Journal of Sport and Exercise Psychology
8	Facing off with scylla and charybdis: a comparison of scalar, partial, and the novel possibility of approximate measurement invariance	van de van de Schoot , R.,Kluytmans, A., Tummers, L. G., Lugtig, P., Hox, J., & Muthén, B.	2013	Frontiers in Psychology
9	Data comparability in the teaching and learning international survey (TALIS) 2008 and 2013	He, J. & K. Kubacka	2015	OECD Publishing
10	BSEM measurement invariance analysis	Muthén, B., &Asparouhov, T.	2013	M*plus* Web Notes

***Note*.** Assigned Number= number assigned by the review author for coding purposes.

Appendix C. PRISMA Flow Chart for the BAMI Systematic Review Citation Process



Appendix D. Wordcloud Showing Terms Used to Describe Prior Informativeness

Appendix E. Examples of SAS Code and M*plus* Code for Data Generations and Models

SAS Code and M*plus* Code for the Data Generations for Population 1: Exact

```
*pop 1 no DIF;
options noxwait xsync;
/***************************************/
/*            Pattern Matrix           */
/***************************************/;
* specify no-DIF groups model;
* step 1: obtain correlation matrix and SDs;
proc iml;
LYG1 = {.8, .6, .5, .6, .8, .5};
PHG1 = {1};
TEG1 = {.36 0 0 0 0 0,
         0 .64 0 0 0 0,
         0 0 .75 0 0 0,
         0 0 0 .64 0 0,
         0 0 0 0 .36 0,
         0 0 0 0 0 .75};
COVG1 = LYG1*PHG1*LYG1` + TEG1;
*print COVG1 ;
     * obtain correlation matrix and SDs from COV;
SG1 = sqrt(diag(covG1));
RG1 = (inv(SG1))*covG1*(inv(SG1));
 create STDG1 from SG1 [colname={y1 y2 y3 y4 y5 y6}];
      append from SG1;
*print SG1 RG1;

* step 2: obtain no-DIF groups pattern matrix;
data A (type = corr);
_TYPE_ ='CORR';
input y1 y2 y3 y4 y5 y6;
cards;
        1 . . . . .
        0.48   1 . . . .
        0.40 0.30   1 . . .
        0.48 0.36 0.30   1 . .
        0.64 0.48 0.40 0.48   1 .
```

```
        0.40 0.30 0.25 0.30 .40   1
;
     *obtain factor pattern matrix for data generation;
proc factor n=6 outstat=facoutG1 noprint;
data patternG1; set facoutG1;
 if _TYPE_ = 'PATTERN';
 drop _TYPE_ _NAME_ ;
run;

/****************************************/
/*           Data Generation            */
/****************************************/;
options nonotes;
*libname ml 'C:\Users\abeer\OneDrive\BSEM18';
%let mc = 100; *number of replications;
%macro manyMI;
     * do loop for groups;
%do grloop = 1 %to 2;
 %if &grloop = 1 %then %do; %let g1 = 8; %let gn = 8; %let pct =
0;%end;
 %if &grloop = 2 %then %do; %let g1 = 20; %let gn = 20; %let pct
= 0; %end;
          * do loop for group size;
     %do gsloop = 1 %to 1;
      %if &gsloop = 1 %then %do; %let gs = 500; %end;

       X mkdir
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\data\gn&gn.gs&gs.
pct&pct.ptnexact-zero\";
               * do loop to generate the given number of
replications;
          %do i=1 %to &mc;
          proc iml;
           n = &g1;
           YB = rannor(J(n, 1, 0));
           zB = YB*.07; *range between -.21 (-3 SD) and .21 (+3
SD); * factor mean variability;
           *print n YB zB;
               /* zB is a set of cluster means */
               %do j = 1 %to &g1;
               use patternG1;
               read all var _NUM_ into FG1;
               FG1 = FG1`;
               use stdG1;
               read all var _NUM_ into STG1;
               YG1 = rannor(J(&gs, 6, 0));
               YG1 = YG1`;
```

```
                zG1 = FG1*YG1;
                zG1 = STG1*zG1;
                zG1 = zG1`;
                cmean = zB[&j,1]; *factor mean variability;
                G1data = zG1 + cmean;
                *print cmean;
                groupid = J(&gs, 1, &j);
                data_G1 = G1data||groupid;
                run;
                * create a SAS dataset per cluster;
                create group&j from data_g1 [colname={y1 y2 y3 y4
y5 y6 group}];
                append from data_g1;
                *print zG1 data_g1;
                %end; *end of g1 loop;

                * stack up groups to create a final dataset;
                data rep&i ;
                     %if &grloop = 1 %then %do;
                      set group1-group8;
                     %end;
                     %if &grloop = 2 %then %do;
                      set group1-group20;
                     %end;

                run;
                proc export data= rep&i
                 outfile =
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\data\gn&gn.gs&gs.
pct&pct. ptnexact-zero\rep&i..dat"
                 dbms = dlm replace ;
                 putnames = no;
                run;
           %end; *end of replication loop;
      %end; *end of group size loop;
%end; *end of group loop;
%mend manyMI; * end of macro;
%manyMI; run;
```

SAS Code to Run Approximate Bayesian Models with Large DIF

```
options noxwait xsync;
*PROC PRINTTO print = "log";

proc iml;
%macro BayesMI(gn,gs,pct,ptn,prior,pr);
```

```
X mkdir
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\Bayes\gn&gn.gs&gs
.pct&pct.ptn&ptn\prior&pr.\";
%do i=1 %to 100;
file
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\Bayes\gn&gn.gs&gs
.pct&pct.ptn&ptn\prior&pr.\rep&i..inp";
put (" data: file is ");
put ("C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\data\")@;put
("gn&gn.")@;put ("gs&gs.")@; put("pct&pct.")@;
put("ptn&ptn.")@;put("\")@;put
("rep")@;put("&i.")@;put(".dat;");
put (" variable: names are y1-y6 group; ");
put ("          usevariables are y1-y6 ; ");
put ("          classes = c (&gn.); ");
put ("KNOWNCLASS = c(group = 1-&gn.);");put;
put (" analysis: type = mixture; ");
put (" estimator = Bayes; ");
put (" processors = 2; ");
put (" model = ALLFREE;");put;
put (" model: %OVERALL%");
put (" f by y1-y6* (1-6);");
put (" [y1-y6] (nu#_1-nu#_6);");
put ("%c#1%");
put ("f@1;");
put ("[f@0];");
put ("MODEL PRIORS:");
put ("DO(1,6) DIFF(nu1_#-nu&gn._#)~N(0,&prior.);");
put (" output: TECH1 TECH8;");
closefile
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\Bayes\gn&gn.gs&gs
.pct&pct.ptn&ptn\prior&pr.\rep&i..inp";
/* CALL MPLUS AND RUN SIMULATION FILES */
X call "C:\Program Files\Mplus\Mplus.exe"
"C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\Bayes\gn&gn.gs&gs
.pct&pct.ptn&ptn\prior&pr.\rep&i..inp"

     "C:\Users\abeer\OneDrive\BSEM18\MI_many_groups\Bayes\gn&gn.
gs&gs.pct&pct.ptn&ptn\prior&pr.\rep&i..out";
%end;
%mend;

%BayesMI (gn=8, gs=500, pct=50, ptn=LG1, prior=.001, pr=001);
%BayesMI (gn=20, gs=500, pct=50, ptn=LG1, prior=.001, pr=001);
%BayesMI (gn=8, gs=500, pct=80, ptn=LG1, prior=.001, pr=001);
%BayesMI (gn=20, gs=500, pct=80, ptn=LG1, prior=.001, pr=001);
```

```
%BayesMI (gn=8, gs=500, pct=50, ptn=LG1, prior=.005, pr=005);
%BayesMI (gn=20, gs=500, pct=50, ptn=LG1, prior=.005, pr=005);
%BayesMI (gn=8, gs=500, pct=80, ptn=LG1, prior=.005, pr=005);
%BayesMI (gn=20, gs=500, pct=80, ptn=LG1, prior=.005, pr=005);


%BayesMI (gn=8, gs=500, pct=50, ptn=LG1, prior=.01, pr=01);
%BayesMI (gn=20, gs=500, pct=50, ptn=LG1, prior=.01, pr=01);
%BayesMI (gn=8, gs=500, pct=80, ptn=LG1, prior=.01, pr=01);
%BayesMI (gn=20, gs=500, pct=80, ptn=LG1, prior=.01, pr=01);


%BayesMI (gn=8, gs=500, pct=50, ptn=LG1, prior=.05, pr=05);
%BayesMI (gn=20, gs=500, pct=50, ptn=LG1, prior=.05, pr=05);
%BayesMI (gn=8, gs=500, pct=80, ptn=LG1, prior=.05, pr=05);
%BayesMI (gn=20, gs=500, pct=80, ptn=LG1, prior=.05, pr=05);


%BayesMI (gn=8, gs=500, pct=50, ptn=LG1, prior=.1, pr=1);
%BayesMI (gn=20, gs=500, pct=50, ptn=LG1, prior=.1, pr=1);
%BayesMI (gn=8, gs=500, pct=80, ptn=LG1, prior=.1, pr=1);
%BayesMI (gn=20, gs=500, pct=80, ptn=LG1, prior=.1, pr=1);
```