A Comparison of Frequentist and Bayesian Approaches for Confirmatory Factor Analysis

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Menglin Xu

Graduate Program in Educational Studies

The Ohio State University

2019

Dissertation Committee

Dr. Richard Lomax, Advisor

Dr. Ann O'Connell, Co-advisor

Dr. Paul De Boeck, Committee Member

Dr. Andrew Hayes, Committee Member

Abstract

Model fit indices within the framework of structural equation models are crucial in evaluating and selecting the most appropriate model to fit data. The performance of fit indices under varying suboptimal conditions requires further investigation. Moreover, with the increasing interest in applying Bayesian method to social sciences data, the comparison of Bayesian estimation and robust maximum likelihood (MLR) estimation methods in evaluating models and estimating parameters is of vital importance. This study aims 1) to investigate the performance of MLR associated model fit indices under various conditions of model misfit, data distribution, and sample sizes; 2) to compare the performance of Bayesian and MLR methods in model fit and parameter estimation based on a confirmatory factor analysis (CFA) model. Data were simulated based on a population CFA model consistent with Curran, West and Finch's (1996) study using R 3.4.0. Simulation conditions include 3 sample sizes (N = 200, 500, 1000), 3 degrees of model misfit (none: RMSEA = 0; mild: RMSEA = .05; moderate: RMSEA = .10), and 3 degrees of data nonnormality (normal: skewness = 0, kurtosis = 0; mild: skewness = 1, kurtosis = 3; moderate: skewness = 2, kurtosis = 7). Model misfit was introduced using Cudeck and Browne's (1992) method through the R package MBESS. Data were fit using the R package lavaan for MLR method and blavaan for Bayesian method. Results show that scaled CFI and scaled TLI are the most robust model fit indices to various

suboptimal conditions; compared to *p* values associated with MLR, PP *p* values associated with the Bayesian method are robust to small sample size and data nonnormality under correctly specified models, less sensitive to models with ignorable degree of misfit, and have sufficient statistical power to reject moderately misspecified models; Bayesian and MLR methods have similar performance in point estimation; MLR method produces more robust standard error estimations. Implications and suggestions for future studies are discussed.

Dedication

Dedicated to those who have provided motivation and support.

Acknowledgments

Thanks to professors for the support in the ideas, programming, and editing of this study; Peers for providing suggestions and friendship with everything; and The Ohio State University for providing the opportunity and academic resources to work towards this degree.

Vita

2010.....B.S. Applied Psychology, Beijing Sport University

2013..... M.Ed. Exercise Psychology, University of Macau

2015 to present .....Graduate Teaching Associate, Department of Educational Studies, The Ohio State University

Publications

Xu, M.L., Leung, S.O. (2016). Bifactor structure for the categorical Chinese Rosenberg Self-Esteem Scale. *The Spanish Journal of Psychology, 19*, 1-11.

Fields of Study

Major Field: Educational Studies

Quantitative Research, Evaluation and Measurement

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Structural equation modeling (SEM) is among the most popular statistical tools in social science, and it models relationship between observed variables and between latent constructs. Confirmatory factor analysis (CFA) is a type of SEM specializing in measurement models which quantifies the relationship between items and the latent constructs an instrument intends to measure. It has high popularity in scale validation and serves as a precursor step for a full SEM model. Major methodological considerations for a CFA model are model fit indices and parameter estimation. Practitioners have long been relying on fit indices for model evaluation and selection, but the behavior of fit indices maintains a thorny issue in methodology. Model fit evaluates how the model implied covariance matrix reproduces the population matrix. Under the framework of maximum likelihood estimation (MLE), a likelihood ratio χ^2 test is conducted through ML fitting function $F_{\text{ML}}(\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ refers to the model implied covariance matrix and $\boldsymbol{\Sigma}$ is the population matrix. Given large sample size, $(N\text{-}1)\, F_{\text{ML}}$ approximates a χ^2 distribution with degree of freedom (*df*) under the null hypothesis. However, it is widely known that the χ^2 test would reject a reasonably fitted model under large sample size, and the χ^2 statistic is inflated with nonnormal variables, users cannot solely rely on the test.

Structural and distributional assumptions need to be considered in examining robustness of SEM (Satorra, 1990). Structural assumption indicates that the fitted model fully captures the inter-relationship among variables in the population, or the covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ based on the fitted model well represents the population. However, as pointed out by MacCallum (2003), models are only artificial approximation of reality under the guidance of theory and experience, so there is no perfect model. He cautioned that conclusions made from simulation studies which generate data from a perfect model would have limited generalizability to empirical settings. It is typical to include model misspecification conditions in simulation designs for studying the behavior of fit indices (e.g., Hu & Bentler, 1998; Fan, Thompson & Wang, 1999). As illustrated by Gerbing and Anderson (2016), the design of model misspecification imposes a challenge to methodological researchers. First, the magnitude and pattern of model misspecification is hard to operationalize, since the magnitude is not solely determined by omitting a nonzero path, but by the whole set of parameters. Second, it is expected to have a prespecified value for fit indices, so that the effect of model misspecification can be comparable and well controlled for in simulation design. Similar concern is shared by Marsh, Hau and Wen (2004) that in typical simulation study designs for misspecification, for example in Hu and Bentler (1998)'s, it is presumed that a model can perfectly reflect the real-world phenomenon, and the misspecification only occurs in certain omitted paths or factor covariances, which is unrealistic for empirical researchers because it is hard to know in advance the exact pattern a misspecification would take.

Such generalizability concern with misspecification design is to some degree relieved by Cudeck and Browne (1992)'s method of manipulating model misfit. Unlike the traditional approach, their method controls the magnitude of misfit through F_{ML} (Σ_0^*, Σ (γ_0)) = c, where Σ is the population model specified by the researcher, e.g., a two-factor CFA model; Σ_0^* denotes the variance-covariance matrix after approximation error is introduced to the population model; c indicates the amount of approximation error desired by the researcher; and γ_0 denotes the sets of parameters minimizing the ML fit function. The method of creating perturbed covariance matrix entails two properties: first, researchers are allowed to specify the magnitude of population misfit in advance, and it is not affected by sample size; second, the perturbed covariance matrix would yield unbiased parameter estimates, thus avoiding the confounding effect of model misfit and other design factors such as nonnormality on parameter estimation. The method is more realistic and the pre-specified approximation error facilitates comparisons across simulation conditions, and therefore is adopted in this study.

Distributional characteristics is another condition that is frequently considered in simulation designs (e.g., Hu, Bentler and Kano, 1992; Curren, West and Finch, 1996; Lei and Lomax, 2005). According to a systematic review by Micceri (1989), more than one half of samples in the educational and psychological field have at least moderate level of skewness. Hence, it is unrealistic to assume empirical data sets to be normally distributed. By reviewing simulation studies involving nonnormality conditions with MLE, West, Finch and Curran (1995) concluded that nonnormality would cause inflated χ^2 statistics, modest downward bias in fit indices such as Tucker and Lewis (1973) Index (TLI) and

Comparative Fit Index (CFI, Bentler, 1990), and moderate to severe downward bias in standard errors. Therefore, nonnormality condition is considered in the simulation design of this study.

Given the known problems with ML χ^2 test, several robust estimators have been proposed such as asymptotic distribution-free methods (ADF; Browne, 1984) and Satorra-Bentler scaled method (SB; Satorra & Bentler, 1994) to accommodate the estimation bias caused by nonnormal variables. Chou, Bentler and Satorra (1991) compared SB, ADF and ML under various conditions of nonnormality and reported that SB performed the best of all. Hu, Bentler and Kano (1992) manipulated seven nonnormality conditions and six sample sizes to compare the performance of ML, generalized least square (GLS), SB, and ADF and made the same conclusion. Similar findings were also shared in Curran, West and Finch's (1996) study. Therefore, ML with Satorra-Bentler scaled χ^2 statistics and standard error (hereafter denoted as MLR) is adopted as the frequentist estimator in this study.

In face of the fact that ML estimation requires assumptions of large sample, multivariate normality, and correct model specification, another estimator of interest in this study is Bayesian. Van, Winter, Ryan, Zondervan-Zwijnenburg and Depaoli (2017) made a systematic summary of Bayesian usage in social science field and found increasing popularity of Bayesian methods. Especially, they reported that in the category of technical and simulation articles, SEM is the second most popular statistical tool, while in the category of applied articles, SEM is the top widely used model, and they foresaw a continuing trend for Bayesian application in SEM in future.

As discussed in Muthén and Asparouhov (2012) and Levy (2016), adoption of Bayesian methods is motivating because it does not rely on large-sample theory or multivariate normality assumption, it allows to incorporate researchers' prior experience and theoretical judgement into parameter estimation through prior specification, and it grants more information of model fit (e.g., posterior predictive *p*-value, denoted as PP *p*-value hereafter) and parameter estimates (e.g., posterior mean, mode, *SD*, and credible interval).

Prior specification is an inseparable part of Bayesian methods. It can be noninformative or informative at various degrees. A noninformative prior distribution conveys little judgement about parameters, for example a uniform distribution for path coefficients; an informative prior distribution implies researchers' own experience with the range or central tendency about parameters, for example a normal distribution with zero mean and unit variance for path coefficients. It is cautioned by Lee and Song (2004) that priors specification has huge impact on posterior distribution, an undesirable prior distribution is even worse than a noninformative prior. In the current study, it is assumed that CFA users do not have specific knowledge of model parameters, and thus noninformative prior distributions are considered.

Based on the previous findings on robustness of SEM (a detailed review is provided in Chapter 2), several improvements can be made. First, while the majority of the simulation studies take a focus on either structural or distributional assumption violations, this study considers both at various levels; second, many previous studies only consider behavior of model fit indices as the outcome variable in simulation design, this

study investigates both fit indices and parameter estimation; third, the design of model misspecification in previous studies is all about either omitting factor covariances or nonzero paths while still assuming a perfect model in the population, limiting its practical value for CFA practitioners, this study incorporates model misfit with a prespecified discrepancy value, which can be inferred from RMSEA in a straightforward manner, facilitating baseline misfit control and interpretation; last, very few studies have compared Bayesian approach with MLR in the context of combined conditions of model fit and nonnormality, this study fills the gap by comparing Bayesian and MLR approaches in terms of model fit and parameter estimation.

The purposes of the study are: 1) To investigate the performance of CFA model fit indices (scaled χ^2 test, CFI, TLI, RMSEA, and SRMR) with robust maximum likelihood estimator (MLR) under varying conditions of misfit and data nonnormality; 2) to compare the performance of MLR and Bayesian estimation in terms of power and parameter estimation.

Chapter 2. Literature Review

In this chapter, concepts of confirmatory factor analysis, model fit indices, and Bayesian method are introduced, prior simulation studies on the performance of maximum likelihood (ML) and Bayesian methods with structural equation modeling are illustrated, and a contrast between ML and Bayesian approaches is discussed.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) seeks to detect underlying constructs that account for correlation or covariance among observed variables. Users are expected to have a priori experience or theoretical rationale to specify the number of latent factors and pattern of loadings. CFA serves as a popular tool for scale validation through dimensionality determination, construct validation, and measurement invariance examination, among others (Brown, 2015). It is posited that responses given to an item is a linear combination of the latent constructs and unique variance. The expression is as follows:

$$\mathbf{y} = \mathbf{\Lambda\eta} + \boldsymbol{\varepsilon}$$

To put it in a covariance matrix form:

$$\mathbf{\Sigma} = \mathbf{\Lambda\Phi\Lambda}' + \mathbf{\Psi}$$

where $\mathbf{y}$ is a p (indicators) × 1 vector of item responses, $\mathbf{\Lambda}$ is a $p \times m$ factor loading matrix, $\boldsymbol{\eta}$ is a $m \times 1$ vector of factors, $\boldsymbol{\varepsilon}$ is a $p \times 1$ vector of measurement error, $\mathbf{\Sigma}$ refers to

the $p \times p$ covariance matrix of responses for all the items involved, $\mathbf{\Phi}$ is a $m \times m$ factor variance-covariance matrix, and $\mathbf{\Psi}$ is a $p \times p$ error variance-covariance matrix.

Model Fit Indices

The χ^2 Test

The χ^2 test (with MLE) indicates the discrepancy between the sample covariance matrix and the hypothesized model. Following Chou and Bentler (1995), the ML fitting function is as follows:

$$F_{ML} = \log|\mathbf{\Sigma(\theta)}| + \text{Trace}(\mathbf{\Sigma(\theta)}^{-1}\mathbf{S}| - \log|\mathbf{S}| - p$$

where $\mathbf{\Sigma(\theta)}$ refers to the model implied matrix, $\mathbf{\theta}$ is the parameter vector minimizing the fit function, $\mathbf{S}$ refers to the sample covariance matrix, and p is the number of variables.

Regarding hypothesis testing, H_0: $\mathbf{\Sigma(\theta) = \Sigma}$ which implies that the model implied covariance matrix perfectly reproduces the population covariance matrix. Under null hypothesis, the following expression approximates a χ^2 distribution.

$$(N\text{-}1)\, F_{ML}$$

where N is the sample size. The χ^2 statistic in relation to the model degrees of freedom (*df*) yields a p value (chi-p) indicating whether the null hypothesis should be rejected or not.

However, the χ^2 test is based on the assumption of large sample, multivariate normality, and correct model specification. Violation of these assumptions would yield biased χ^2 statistics. Additionally, it tests exact fit, which is not of practical use because it is known that a best fit model is only an approximation to population (MacCallum, 2003). Other fit indices have been developed to capture various aspects of model performance.

Incremental Fit Indices

Incremental fit indices quantify the improvement of a hypothesized model against a null model, which is a restricted model assuming no relationship among variables. Comparative Fit Index (CFI; Bentler, 1990) and Tucker-Lewis Index (TLl; Tucker and Lewis, 1973) are included in this study. Following Schumacker and Lomax (2016, *pp*.106-143), CFI and TLI are expressed as follows:

$$\text{CFI} = 1 - \frac{\max(0,\ \chi^2_{model} - df_{model})}{\max(0,\ \chi^2_{null} - df_{null})}$$

$$\text{TLI} = \frac{\frac{\chi^2_{null}}{df_{null}} - \frac{\chi^2_{model}}{df_{model}}}{\frac{\chi^2_{null}}{df_{null}} - 1}$$

A value closer to 1 represents better fit, a value larger than .95 was considered satisfactory for both CFI and TLI (Hu & Bentler, 1999).

Absolute Fit Indices

Root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980) and Standardized Root-mean square residual (SRMR; Bentler, 1995) are included in this study. The calculations are as follows:

$$\text{RMSEA} = \sqrt{\frac{\max(0,\ \chi^2_{model} - df_{model})}{(N-1)df_{model}}}$$

$$\text{SRMR} = \sqrt{\sum_j \sum_{k \le j} \frac{r_{jk}^2}{p(p+1)/2}}$$

where r_{jk}^2 refers to the squared elementwise difference between the sample and model implied correlation matrix, and p is the number of variables. A value less than 0.05 or

0.06 is considered as satisfactory for RMSEA, and a value less than 0.08 is considered as decent for SRMR (Hu & Bentler, 1999).

Prior Simulation Studies on ML/MLR Performance

Curran, West, Finch (1996) conducted a Monte Carlo simulation study to investigate the effects of nonnormality, sample size, and model misspecification on behaviors of the χ^2 test estimated with ML, and Satorra-Bentler (SB) ML. The data generation model is a three-factor nine-indicator CFA model. Three levels of univariate skewness and kurtosis were considered representing normal, moderate nonnormal, and extremely nonnormal distribution of variables. Model misspecification refers to either adding a path which is zero in the population model, or omitting a nonzero cross loading. Four levels of sample size were considered ranging from 100 to 1000. Results showed that under the condition of correct model specification and normal distribution, both ML and SB-ML yield unbiased χ^2 estimates regardless of sample size, and ML- χ^2 is inflated as nonnormality increases; under the condition of model misspecification, as nonnormality increases, ML- χ^2 has upward bias, while SB-ML has downward bias and thus its power to detect model misspecification decreases.

Nevitt and Hancock (2000) conducted a simulation study to explore effects of model specification and nonnormality on performance of ML-based and SB-based RMSEA. Their simulation design is very similar to Curran et al.'s (1996) including four levels of sample size, two levels of misspecification, and three levels of nonnormality. Their results showed that when the model is correctly specified, as nonnormality increases, both ML-based and SB-scaled RMSEA have increasing average value, with

SB-scaled RMSEA displaying better Type I error control; when the model is misspecified and data is normal, ML-based RMSEA has power close to the nominal level only when sample size is 1000, and SB-scaled RMSEA has similar behavior; as nonnormality increases, ML-based RMSEA increases while SB-scaled RMSEA decreases. All these findings are referring to the close fit: reject H_0 when RMSEA > .05.

Hu and Bentler (1998, 1999) investigated sensitivity of fit indices to model misspecification under systematically varied conditions of nonnormality and sample sizes. The data generation model is a three-factor fifteen-indicator CFA model. Model misspecification is defined by either covariance (e.g., collapsing two factors into one) or loading (e.g., omitting a cross-loading) misspecification. Seven combined conditions of skewness and kurtosis were added to factor scores, measurement errors, or both. Sample sizes ranged from 150 to 5000. Their results revealed that CFI, TLI, RMSEA belong to a cluster and are more sensitive to loading misspecification, while SRMR behaves the least similar to others and is more sensitive to covariance misspecification. Therefore, they suggested a two-index strategy for fit indices reporting, ML-based CFI, TLI, RMSEA together with SRMR. In addition, they proposed cutoffs for the indices, which are 0.95 for CFI and TLI, 0.06 or 0.05 for RMSEA, and 0.08 for SRMR. The findings have exerted widespread influence in guiding SEM practitioners.

Despite the popularity of their study, the generalizability to models in other settings have been called into question. As pointed out by Marsh, Hau and Wen (2004), cautions should be made in applying the cutoffs to research by SEM users. First, the magnitude of misspecifications in Hu and Bentler's (1998, 1999) study falls into an

acceptably misspecified range according to the criteria per their own suggestion (e.g., TLI > 0.95). In addition, there is a paradox in the behavior of indices, with misspecified models, small sample size would lead to higher false rejection rate of a correct model and false acceptance of a misspecified model, while large sample would lead to zero rejection rate of a correct model and a 100% false acceptance rate for a misspecified model. Moreover, the indices other than SRMR have higher rejection rate for loading misspecification than for covariance misspecification, while the reverse pattern is found for SRMR. Marsh et al. (2004) further cautioned that not only the magnitude, but also the pattern of misspecification is limited in its representativeness in real life, because it assumes a true model perfectly representing the population while misspecification is narrowly defined as either omitted factor covariances or cross loadings.

Fan and Sivo (2005) also raised the generalizability concern with Hu and Bentler (1998, 1999)'s conclusions in the sense that the magnitude and type of model misspecification are confounded with loading misspecification having larger magnitude. They replicated the design and adjusted population parameters so that the magnitude of misspecification, as operationally defined as the statistical power of χ^2 test to reject an incorrect model, can be comparable across models. When correlating fit indices, it was found that SRMR is less correlated than the others, but a subsequent exploratory factor analysis suggests a clear one-factor solution, indicating that SRMR is not behaving exclusively from others. Their ANOVA results also suggested that SRMR is more sensitive to covariance misspecification while the other indices more sensitive to loading misspecification. However, once the factor covariance misspecification was adjusted to

be fixed to one, SRMR does not have differential sensitivity to the two misspecification types, casting doubts on the two-index strategy.

Lei and Lomax (2005) investigated effects of nonnormality on performance of fit indices and parameter estimates of SEM. The data generation model is a SEM model with one exogenous and two endogenous latent variables, each with two indicators. Sample sizes range from 100 to 1000. Seven combined levels of univariate nonnormality with manifest variables were considered, a skewness around 0.3 and kurtosis around 1.0 were specified for slight nonnormality, a skewness > 0.7 and kurtosis > 3.5 were specified for severe nonnormality. Results showed that nonnormality does not significantly affect standard errors across samples sizes; for both exogenous and endogenous variables, nonnormality exerts more important impact than sample size on bias of parameter estimates, and sample sizes have larger effect on endogenous variables than on exogenous variables; small sample sizes (less than 500) significantly affect all fit indices including χ^2, TLI, and CFI; large sample size and nonnormality significantly affects χ^2 statistics only, revealing that the χ^2 statistic is lacking robustness.

Yang and Liang (2013) studied effects of nonnormality and model misspecification on model fit indices and parameter estimates for two-factor ten-indicator CFA models. Nine combined levels of nonnormality on factor scores and errors were considered, skewness = 0.75 and kurtosis = 1.75 were specified for second degree of nonnormality, skewness = 1.25 and kurtosis = 3.75 were specified for third degree of nonnormality. Model misspecification is defined as either omitting a factor covariance, or omitting a nonzero cross loading. Results showed that for misspecified models, CFI and

RMSEA with ML/MLR were insensitive to main effects of sample size and nonnormality and two-way interactions between them. In terms of parameter estimation, ML/MLR estimates were robust to nonnormality at the moderate level.

Xia, Yung and Zhang (2016) examined the performance of robust χ^2 tests and standard error estimation of a CFA model under varying conditions of sample size, data nonnormality, and model misfit. They reported that under no model misfit, robust χ^2 rejection rates are inflated when $N = 200$ or smaller, and become closer to .05 when sample size reaches 500; under mild model misfit, robust χ^2 over-reject the model; under higher degrees of model misfit, robust χ^2 have sufficient statistical power to reject the model even when $N = 200$. In terms of standard error estimation, regardless of degrees of model misfit, it decreases with sample size and increases with data nonnormality. The bias in standard error estimation increases slightly with data nonnormality, but with a negligible magnitude when sample size reaches 200.

Bayesian Estimation

Bayesian analysis produces posterior distribution for parameters through Bayes theorem as follows.

$$p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

where $p(\boldsymbol{\theta}|y)$ refers to the probability to observe parameter vector $\boldsymbol{\theta}$ given the available data, $p(y|\boldsymbol{\theta})$ refers to likelihood of data, and $p(\boldsymbol{\theta})$ is prior distribution for parameters. The theorem shows that the posterior distribution of parameters is proportional to the product of likelihood of data and prior information. It indicates that the information in data is augmented by priors and then yields the posterior distribution.

The choice of priors is influential to posterior estimates. When researchers have no prior concept or experience with parameter estimates, noninformative priors such as a uniform distribution $U(0, 10^6)$ can be used for a non-negative regression coefficient. If researchers have prior experience about model parameters, or gain information from articles in the same field, informative priors can be specified such as a normal distribution $N(0.5, 1)$ for a regression effect. Conjugate priors are commonly used in available software (e.g., blavaan package by Merkel & Rosseel, 2015). Data likelihood updated by a conjugate prior would produce the posterior distribution which is in the same distributional family with priors, easing computational burden. Based on the CFA model in this study, the noninformative default priors adopted in the R package blavaan are as follows for a CFA based covariance matrix:

$$\Sigma = \Lambda\Phi\Lambda' + \psi$$

Priors are:

$$\Lambda \sim N(0, 100)$$

$$\Phi \sim \text{Inverse Wishart}(\text{I}, 4)$$

$$\psi \sim \text{Inverse Gamma}\ (1, 0.5)$$

All the factor loadings follow a normal distribution with mean of zero and variance of 100. The variance-covariance matrix of the latent factors follows an inverse Wishart distribution with the scale matrix being a 3*3 identity matrix and degrees of freedom being 4. Each error variance follows an inverse Gamma distribution with shape parameter of 1 and scale parameter of 0.5. According to Merkel and Rosseel (2015), the default priors are conjugate and proper.

Markov Chain Monte Carlo (MCMC)

Bayesian analysis is based on Markov chain Monte Carlo (MCMC) sampling which draws samples from posterior distribution of parameters rather than calculating high-dimensional numerical integration. In blavaan, Gibbs sampling is used. Following Kaplan and Depaoli (2015), the idea of Gibbs sampler is as follows:

Sample θ_1^s from $p(\theta_1|\theta_2^{s-1}, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, y)$, then

Sample θ_2^s from $p(\theta_2|\theta_1^s, \theta_3^{s-1}, \ldots, \theta_q^{s-1}, y)$

Sample θ_q^s from $p(\theta_q|\theta_1^s, \theta_2^s, \ldots, \theta_{q-1}^s, y)$

Starting from a set of initial values, the Gibbs algorithm samples a first parameter θ_1^s from the conditional distribution of θ_1 given the data and all the other parameters, then draws a second parameter θ_2^s conditional on the other parameters in the previous draw, all the way through θ_q^s where q denotes the number of parameters, until the pre-specified number of iterations is reached.

Convergence Diagnosis

Convergence is diagnosed by a potential scale reduction factor (PSR) (Gelman & Rubin, 1992). It requires at least two Markov chains, and it implies the ratio of between-chain variance to within-chain variance. The ratio shall be close to 1.0 if the chains are mixing well to a stationary status. The current study adopts two chains, and specifies that any parameter estimates exceeding PSR of 1.2 is considered to have a serious convergence problem, and any replicate having a convergence concern is discarded and not considered for further analysis. Convergence diagnosis is also aided by checking trace plots and autocorrelation plots. Trace plots tell whether the two chains are mixing

well and proceeding smoothly as the number of iterations increases; autocorrelation plots tell whether draws at a certain lag, say every 10th, is relatively independent from the previous draws, and a near-zero correlation between draws symbolizes good convergence.

Posterior Predictive p-value

Model fit for Bayesian estimation is represented by the posterior predictive *p*-value (PP *p*-value) in this study. It is an exact fit index reflecting how close the replicated data is from the raw data. Following Gelman, Carlin, Stern and Rubin (2004), pthe osterior predictive distribution is expressed as:

$$p(y^{rep}|y) = \int p(y^{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|y)d\boldsymbol{\theta}$$

where y^{rep} refers to replicated data, $p(y^{rep}|\boldsymbol{\theta})$ refers to the probability of y^{rep} given the parameter vector $\boldsymbol{\theta}$, which is drawn from the posterior distribution denoted by $p(\boldsymbol{\theta}|y)$. After obtaining a sample of y^{rep} , a test statistic shall be defined as the discrepancy measure, which is a χ^2 statistic in this study. The χ^2 statistic is calculated for the raw data and the replicated data sets, and PP *p*-value is calculated as:

$$\text{PP } p\text{-value} = p(\chi^2(y^{rep}) \geq \chi^2(y)|y)$$

which means the proportion of replications where χ^2 statistics based on y^{rep} is larger than or equal to that of raw data.

The closer a PP *p*-value is to 0.5, the better the model fit. It is worth mentioning that PP *p*-value has quite different interpretations from the *p*-value associated with χ^2 test in MLR. Although both are based on χ^2 statistics, the *p*-value in MLR implies that under the null hypothesis of χ^2 central distribution, the probability to observe values falling into

the critical region, while PP *p*-value within Bayesian framework implies that how well the replicated data mimic the raw data. For the sake of comparing rejection rate of *p*-value both with MLR and Bayesian approaches, a cut-off point of .05 is specified for PP *p*-value. Asparouhov and Muthen (2010) did a series of simulation studies to explore the performance of PP *p*-value and argued that a cut-off of .05 is reasonable. Therefore, analogous to MLR estimator, a PP *p*-value < .05 is considered as successfully detecting the model misfit given there is misfit or as wrongly rejecting the model given a correctly specified model. In the blavaan package, other model fit indices are available such as Deviance Information Criterion (DIC), Widely Applicable Information Criterion (WAIC), given that in the current study, no model comparisons or model selections are concerned, these indices are thus not relevant to this study.

MLR vs Bayesian

Comparing MLR with Bayesian estimation, first, MLR treats data as random and parameters as fixed, it seeks to find the maximum value for the likelihood surface; in contrast, Bayesian treats data as fixed and parameters as random, it seeks to produce the posterior distribution for parameters, and users are free to obtain mean, mode or other descriptive statistics based on the whole distribution. Second, in MLR, the confidence interval (CI) is interpreted as out of a huge number of samples for CI of a certain parameter, 95% of the CIs would contain the true value; while in Bayesian, the posterior distribution is an empirical distribution for a parameter, therefore the credible interval can be inferred from percentiles, and its interpretation is straight-forward: the probability for the true value to fall in the credible interval is .95. Third, the two approaches can give rise

to similar results when sample size is large and the prior is a uniform distribution, then point estimates from the maximum likelihood method is similar to the posterior mean in Bayesian approach, and the confidence interval is similar to the credible interval.

Prior Studies Comparing Frequentist vs Bayesian

Lee and Song (2004) compared Bayesian versus ML estimation for simulated CFA and SEM models with small sample sizes. They used priors obtained from initial runs adopting noninformative prior distribution with Bayesian method. Sample sizes were manipulated varying from 2 times to 5 times the number of parameters to be estimated, and models with varying magnitude of parameters were tested. They found that under normal data distribution, the PP *p*-value has accurate (close to .5) and stable performance across sample size conditions, while the chi-square test under ML approach has an inflated rejection rate for correctly specified models, and the chi-square statistic is deviating from the chi-square distribution; in terms of parameter estimation, Bayesian approach yields accurate results with acceptable root mean square (RMS), while ML approach yields biased estimates with large RMS, showing that Bayesian performs better than ML both in goodness-of-fit and parameter estimation under small sample sizes. They further reported that under nonnormal data distribution, Bayesian performs less stable than in normal condition, and the estimates of variances and covariances are noticeably worse.

Muthen and Asparouhov (2012) compared performance of MLE and Bayesian methods (with noninformative priors) based on simulated CFA models. They manipulated sample sizes, and model misspecification defined by various magnitudes of

cross-loadings or residual correlations. They reported that when the model is correctly specified, PP *p*-values rejection rates are close to the level of .05, while ML-*p* values have slight inflation; when the model is mildly misspecified, ML-*p* values are over-sensitive to the misspecification; when the degree of misspecification is higher, ML-*p* values have higher rejection rates than PP *p*-values, but both display sufficient statistical power to reject the model as long as sample size reaches 200. These results agreed with each other when the model misspecification was defined by cross-loadings and by residual correlations respectively. In terms of parameter estimation, they reported that the two estimation methods yielded highly comparable parameter accuracy and thus no preference was made.

Liang and Yang (2016) compared MLR and Bayesian estimation (with non-informative priors) based on CFA and bi-factor models with varying model complexity. They manipulated sample size, nonnormality of data, and model misspecification defined by collapsing latent factors. They found that under the condition of normal data and relatively small sizes, the Bayesian method has lower statistical power to detect a misspecified model and lower Type I error when the model is correct; moreover, such pattern is offset by increasing degrees of nonnormality. In addition, they found the utility of the Bayesian method to detect nonnormality when model is correct even with small sample size, while the chi-square test in MLR tends to reject correct models too often for models with large number of variables (e.g., 16). They also reported that Bayesian estimation yields comparable parameter estimates and standard errors with MLR across

simulation conditions, and suggested the usage of the Bayesian method in parameter estimation under nonnormality conditions.

Chapter 3. Method

A Monte Carlo simulation study is conducted for evaluating MLR and Bayesian estimators for a CFA model. In this chapter, the simulation design, data generation procedures, data analysis, and evaluation criteria are illustrated.

Design of the Simulation Study

Design factors include 3 sample sizes × 3 levels of model misfit × 3 degrees of data nonnormality, all factors being fully crossed, yielding 27 conditions. In each condition, 200 replications are performed, thus generating 27*200 = 5400 data sets. Each data set is then fit to a CFA model using MLR and Bayesian estimators repeatedly. In the Bayesian analysis, noninformative priors are involved which suits the situation when CFA users have no prior experience for the parameters.

Population Model

The CFA model used to generate data is based on Curran, West and Finch (1996)'s study. As shown in Figure 1, it is a 3-factor model with three indicators loaded on each factor. All the factor loadings are set to 0.70, error variances are 0.51, resulting in unit variance in each indicator. The factor variances are 1.0, and the inter-factor correlations are 0.30.

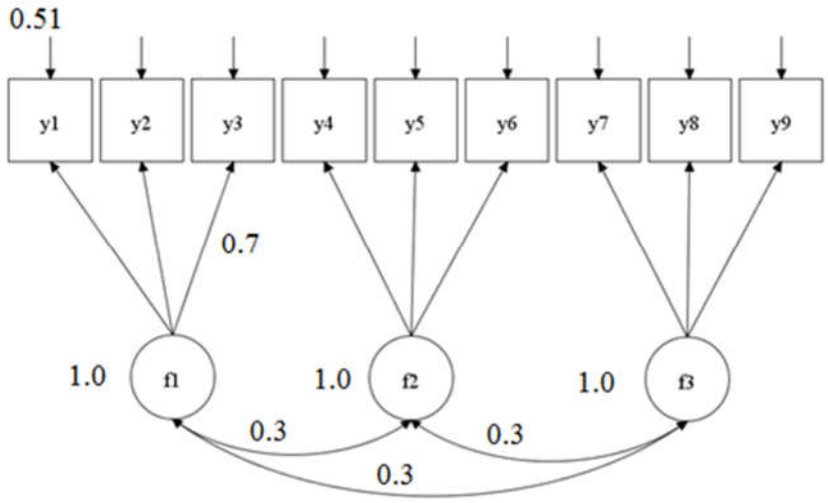


Figure 1. Three-Factor CFA Model with Data Generation Parameters

Sample Sizes

Three levels of sample size are manipulated to be 200, 500, and 1000 indicating small, medium, and large samples. These levels are commonly adopted in simulation research (e.g., Curran et al., 1996).

Data Nonnormality

Three levels of data distribution are considered. Normal: skewness = 0, kurtosis = 0; mild nonnormality: skewness = 1, kurtosis = 3; moderate nonnormality: skewness = 2, kurtosis = 7. All the distributional characteristics target at the observed variables. These distributions are chosen because they reflect common distribution in the field of applied psychology (Micceri, 1989).

Model Misfit

The current study adopted Cudeck and Browne's (1992) method to specify model misfit. Three levels of misfit are considered, no misfit ($\delta = 0$), mild misfit ($\delta = 0.06$), and

moderate misfit (δ = 0.24). These values are chosen because they correspond to RMSEA of 0, 0.05, and 0.1, indicating perfect fit, good fit, and poor fit (Browne & Cudeck, 1993). No extremely bad misfit is considered here because for SEM practitioners, no further attention is expected to be paid to a model if RMSEA is larger than 0.10.

Data Generation Procedure

All the data were generated using R 3.4.0 (R Core Team, 2017). Based on the population model as depicted in Figure 1, the population variance-covariance matrix was generated, denoted as Σ. Then, perturbed covariance matrix was generated using Cudeck and Browne's (1992) method. Explicitly, the R package MBESS (Kelley, 2018) was used to incorporate approximation error into Σ using Lai's (2018) function Sigma.2.SigmaStar(). The input is model specification and the desired error (the value *c*). The value *c* is derived from RMSEA thorough the expression $c = \text{RMSEA}^2 * df$ where *df* refers to the degrees of freedom of the population model (*df* = 24). Based on a simulated large sample of 100000, the three levels of misfit correspond to model fit indices of [perfect fit: CFI = 1.000, TLI = 1.000, RMSEA = .002], [good fit: CFI = .972, TLI = .958, RMSEA = .051], [poor fit: CFI = .896, TLI = .845, RMSEA = .101], revealing a good recovery of model fit indices. The variance-covariance matrix after incorporated with increasing levels of approximation error are hereafter denoted as Σ_0, Σ_1, and Σ_2.

Data nonnormality was introduced afterwards. Multivariate normal data was generated in R based on the three variance-covariance matrices. For each misfit condition, three levels of nonnormality were incorporated using Fleishman's (1978)

method. It manipulates univariate skewness and kurtosis through the following transformation expression:

$$X = a + bZ + cZ^2 + dZ^3$$

Where X is the transformed data, Z is the raw data which are normally distributed, *a, b, c* and *d* are the coefficients derived from the desired skewness and kurtosis. The coefficients are available by checking Fleishman table (Fleishman, 1978; Fan, Felsovalyi, Sivo & Keenan, 2002). Table 1 and Table 2 display the covariance matrices for data generation models under selected simulation conditions.

Table 1 *Covariance Matrices for Data Generation Models with Three Degrees of Misfit*

	V1	V2	V3	V4	V5	V6	V7	V8	V9
				Σ_0 : no misfit					
V1	1								
V2	0.49	1							
V3	0.49	0.49	1						
V4	0.147	0.147	0.147	1					
V5	0.147	0.147	0.147	0.49	1				
V6	0.147	0.147	0.147	0.49	0.49	1			
V7	0.147	0.147	0.147	0.147	0.147	0.147	1		
V8	0.147	0.147	0.147	0.147	0.147	0.147	0.49	1	
V9	0.147	0.147	0.147	0.147	0.147	0.147	0.49	0.49	1
				Σ_1 : mild misfit					
V1	1								
V2	0.516	1							
V3	0.491	0.463	1						
V4	0.078	0.143	0.203	1					
V5	0.083	0.148	0.208	0.499	1				
V6	0.091	0.155	0.215	0.49	0.48	1			
V7	0.076	0.141	0.201	0.101	0.141	0.176	1		
V8	0.084	0.148	0.208	0.108	0.149	0.183	0.493	1	
V9	0.092	0.156	0.217	0.116	0.157	0.192	0.49	0.487	1
				Σ_2 : moderate misfit					
V1	1								
V2	0.541	1							
V3	0.491	0.438	1						
V4	0.014	0.138	0.255	1					
V5	0.024	0.148	0.264	0.508	1				
V6	0.038	0.163	0.279	0.491	0.472	1			
V7	0.01	0.135	0.251	0.057	0.136	0.203	1		
V8	0.025	0.149	0.266	0.072	0.151	0.217	0.496	1	
V9	0.04	0.165	0.281	0.087	0.166	0.233	0.49	0.484	1

Table 2 *Covariance Matrices for Data Generation Models with Three Degrees of Misfit Plus Moderate Nonnormality*

	V1	V2	V3	V4	V5	V6	V7	V8	V9
No misfit + moderate nonnormality									
V1	0.99								
V2	0.302	0.939							
V3	0.297	0.219	1.015						
V4	0.056	0.103	0.056	1.005					
V5	0.033	0.074	0.089	0.341	1.019				
V6	0.017	0.104	0.073	0.339	0.345	1.077			
V7	0.061	0.027	0.025	0.041	0.08	0.094	0.846		
V8	0.091	0.033	0.057	0.082	0.084	0.11	0.308	0.982	
V9	0.012	0.048	0.045	0.057	0.034	0.069	0.241	0.288	1.028
Mild misfit + moderate nonnormality									
V1	0.989								
V2	0.269	0.846							
V3	0.288	0.287	0.936						
V4	-0.057	0.089	0.054	1.062					
V5	0.006	0.109	0.097	0.333	1.092				
V6	-0.007	0.069	0.062	0.342	0.321	0.963			
V7	0	0.02	0.109	0.065	0.063	0.056	1.099		
V8	-0.031	0.046	0.083	0.091	0.114	0.074	0.374	1.083	
V9	0.02	0.074	0.035	0.034	0.028	0.089	0.278	0.377	1.003
Moderate misfit + moderate nonnormality									
V1	1.039								
V2	0.321	1.058							
V3	0.315	0.289	1.058						
V4	-0.002	0.102	0.176	1.059					
V5	0.037	0.119	0.207	0.416	1.047				
V6	-0.053	0.073	0.138	0.344	0.354	0.977			
V7	0.033	0.06	0.095	0.081	0.092	0.13	1.182		
V8	0.021	0.099	0.181	0.092	0.128	0.177	0.389	0.989	
V9	0.015	0.052	0.085	0.1	0.095	0.153	0.425	0.324	0.962

Model Estimation

For the combined 27 conditions, 200 replicates were performed. The data was then fit to a same CFA model with the population model using the R package lavaan (Rosseel, 2012) for MLR and blavaan (Merkle & Rosseel, 2015) for Bayesian estimation. In lavaan, estimator = "MLR" is used. In blavaan, several trials were initially made to gain an empirical understanding of the number of iterations needed for convergence. It is tested that with burnin = 5000 and number of iterations = 15000, model convergence was reached even for the largest degree of nonnormality and misfit. Therefore, the settings were adopted for every run with two chains.

Results Saving

The outcomes saved include, for each replication per condition, standardized parameter estimates together with their standard errors (*SE*) for the CFA model, model fit indices of scaled χ^2, p value for scaled χ^2, RMSEA, CFI, TLI (all based on scaled χ^2), and SRMR for MLR estimator, since these are among the most commonly reported indices in published articles, and are consistent with those reported in Mplus output, which facilitates comparisons with Mplus results popularly adopted in journal articles in social science. For Bayesian estimator, PP p-value, parameter estimates with standard error were saved. Model convergence constitutes a crucial concern for both MLR and Bayesian, variables indicating convergence were also programmed in R and saved in results. Additionally, since inadmissible cases would distort the interpretation of estimation results, inadmissible issue was defined as either negative factor variance or

error variance of indicators, programmed in R, and saved. Only replicates without convergence and inadmissibility concerns were considered further.

Data Analysis

Based on the saved replicates, those with nonconvergence or inadmissible issues were recorded and discarded from further analysis. First, the percentages of nonconvergence or inadmissible cases were reported for each crossed condition to gain an initial insight to the estimation performance. Second, for MLR estimator, ANOVA analysis was conducted using SPSS 24.0 to examine the effects of design factors of sample size, degree of misfit and nonnormality on each fit index separately to gain understanding of sensitivity of indices to various suboptimal conditions. Main effects and two-way interactions were considered in ANOVA models. For Bayesian analysis, the ANOVA model was fit to examine the sensitivity of PP *p*-value to conditions. Third, for both MLR and Bayesian estimators, mean values of point estimates and *SE*s were calculated for later use of parameter estimation performance evaluation. Fourth, in order to compare MLR vs Bayesian in terms of statistical power in rejecting misfit models, rejection rates (α = .05) of scaled χ^2 associated *p*-value and PP *p*-value were calculated across conditions.

Evaluation Criteria

In terms of sensitivity of model fit indices, η^2 is calculated using the formula $\frac{SS_{effect}}{SS_{total}}$, where SS_{effect} denotes the Type III sum of square for a main effect or an interaction effect in the ANOVA model, and SS_{total} denotes the corrected total effect provided in the ANOVA table. Eta squared (η^2) is used because if its clear and

straightforward interpretation of effect size, that is, the proportion of variation in the outcome accounted for by the main effect or interaction effect, which is equivalent to R^2. A value of .10 for η^2 would be considered as an adequate amount of variance explained by a predictor (Cohen, 1988).

In regard to parameter estimation performance, relative bias (RB) serve as the criterion for both point estimates and standard error estimates. The calculation is as follows:

$$RB = \frac{|\hat{\theta} - \theta|}{\theta}$$

Where $\hat{\theta}$ denotes the mean estimates over replicates, θ denotes the population value; regarding *SE* estimates, θ is given by the *SD* of the corresponding point estimates. The cutoff for *RB* is .10 for point estimates and standard error estimates (Hoogland and Boomsma, 1998).

With respect to statistical power to reject misfit models, rejection rates of scaled χ^2 associated *p*-value or PP *p*-value exceeding 80% are considered appropriate.

Chapter 4. Results

In this chapter, I began with displaying the percentage of solutions that had model nonconvergence or inadmissible solution issues by design factors. After removing the inadmissible solutions, a series of factorial ANOVA models were performed to examine the effects of design factors (sample size, model misfit, and data nonnormality) and their interactions on each model fit index associated with MLR. The following sections present the results for the differences between MLR and Bayesian estimation methods in terms of power, and parameter estimations. Power (or Type I error under correctly specified model) was evaluated by the rejection rates of *p* values associated with the two methods. Parameter estimations were discussed based on relative biases of loading and inter-factor correlation estimations, standard errors, and relative biases of standard error estimations. Descriptive plots, factorial ANOVA models for outcomes associated with each estimation method, and mixed-design ANOVA models accommodating both between-subjects effects and within-subjects effects were adopted to examine the differences between MLR and Bayesian methods.

Nonconvergence and Inadmissible Solutions

Table 3 displays the percentage of replicates with nonconvergence issues and inadmissible solutions. For the column names, *N* refers to sample size, cov refers to model misfit, and dist refers to data distribution. The notations are hereafter adopted in

the subsequent tables. The highest percentage was associated with the combined condition of small sample size ($N = 200$), moderate nonnormality, and moderate misfit. When sample size was larger ($N = 500$ or 1000), no convergence issue or inadmissible replicates were found. It reveals that small sample size, higher degree of misfit and nonnormality are more likely to have nonconvergence and inadmissibility problems, which is consistent with assumptions with maximum likelihood estimation. Overall, 39 (0.72%) of the simulated samples were removed from further analyses.

Table 3 *Percentage (%) of Nonconvergence and Inadmissibility*

		cov		
N	dist	a	b	c
200	1	0	0	0.5
	2	0	0	1.5
	3	3	5	9.5
500	1	0	0	0
	2	0	0	0
	3	0	0	0
1000	1	0	0	0
	2	0	0	0
	3	0	0	0

Note. *N*: sample size; dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal; cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Sensitivity of Model Fit Indices

Descriptive Statistics

Table 4 shows the sample means for the model fit indices associated with MLR estimator by design factors. When there was no misfit (cov = a), all five indices indicate satisfactory fit to the data, namely, χ^2 close to the degree of freedom of the CFA model (*df* = 24), scaled CFI and scaled TLI were above 0.95, and both scaled RMSEA and SRMR were less than .05. As the degree of misfit increased, the indices showed decreasing model fit, which is consistent with the expectations from the simulation design. Sample size seemed to not affect model fit indices remarkably, with the exception of scaled χ^2 which increased with sample sizes under misspecified model conditions. For scaled χ^2, scaled RMSEA, and SRMR, higher degree of nonnormality appeared to be related to poorer model fit under correct model specification, while opposite pattern was observed under model misfit, suggesting the need to consider interaction effects of the design factors on the performance of model fit indices in the subsequent ANOVA models.

Table 4 *Sample Means of Model Fit Indices across Design Factors*

N	dist	Scaled χ^2	Scaled CFI	Scaled TLI	Scaled RMSEA	SRMR
			cov = a (no misfit)			
200	1 (Normal)	24.741	0.993	0.998	0.016	0.035
	2 (Mild NN)	26.467	0.989	0.991	0.021	0.037
	3 (Moderate NN)	28.779	0.957	0.954	0.026	0.043
500	1 (Normal)	24.623	0.997	0.999	0.01	0.022
	2 (Mild NN)	24.921	0.996	0.999	0.011	0.024
	3 (Moderate NN)	25.574	0.987	0.993	0.012	0.027
1000	1 (Normal)	24.09	0.999	1	0.007	0.016
	2 (Mild NN)	24.009	0.999	1	0.007	0.016
	3 (Moderate NN)	24.957	0.995	0.998	0.008	0.019
			cov = b (mild misfit)			
200	1 (Normal)	38.133	0.967	0.952	0.051	0.048
	2 (Mild NN)	36.635	0.967	0.952	0.046	0.047
	3 (Moderate NN)	30.583	0.946	0.934	0.031	0.045
500	1 (Normal)	53.941	0.973	0.959	0.049	0.039
	2 (Mild NN)	52.966	0.969	0.954	0.047	0.039
	3 (Moderate NN)	32.739	0.971	0.959	0.023	0.032
1000	1 (Normal)	84.712	0.972	0.958	0.05	0.036
	2 (Mild NN)	76.834	0.971	0.957	0.046	0.035
	3 (Moderate NN)	37.147	0.978	0.968	0.021	0.025
			cov = c (moderate misfit)			
200	1 (Normal)	76.605	0.891	0.836	0.103	0.072
	2 (Mild NN)	73.726	0.882	0.823	0.1	0.071
	3 (Moderate NN)	41.499	0.88	0.823	0.053	0.052
500	1 (Normal)	159.165	0.885	0.828	0.104	0.067
	2 (Mild NN)	137.784	0.888	0.83	0.097	0.065
	3 (Moderate NN)	51.707	0.886	0.878	0.045	0.041
1000	1 (Normal)	281.231	0.891	0.837	0.103	0.065
	2 (Mild NN)	246.506	0.888	0.832	0.096	0.063
	3 (Moderate NN)	73.088	0.923	0.885	0.044	0.037

Note. NN: nonnormal.

Factorial ANOVA Results

With the aim of examining the effects of the design factors on model fit indices, factorial ANOVA models were fitted for each fit index separately. Main effects and interactions were included in the models to enable a better understanding of the effects. Practical significance was indicated by η^2, which is calculated using the formula $\frac{SS_{effect}}{SS_{total}}$, as was illustrated in Chapter 3. The statistics represents the proportion of variance in the outcome explained by a predictor. Effects with η^2 larger than or close to .10 were considered as having adequate contribution to the model. Table 5 presents the ANOVA results for scaled χ^2. According to η^2, all the three design factors had meaningful impacts on the outcome, and model misfit had the strongest influence, accounting for 37.3% of the variation in scaled χ^2. Tukey's Post Hoc multiple comparison results for the three factors are displayed in Table 6 through Table 8. It can be told from the tables that scaled χ^2 inflated with larger sample size and higher degree of misfit, and decreased with nonnormality.

It is worth noting that the interaction between model misfit and sample size, and the interaction between misfit and distribution had practical significance around 0.1 as indicated by η^2, which suggests that about 10% of the variation in the outcome were explained by the interactions each. In the presence of significant interactions, it is more important to understand the effect of a design factor within each level of another factor involved. Therefore, interaction plots based on marginal means are provided. As shown in the upper panel of Figure 2, when the model was correctly specified, there were no differences across sample sizes. The sample size differences became the largest when the

model was moderately misspecified, with larger sample size yielding more inflated Scaled χ^2. As shown in the lower panel of Figure 2, when there was no misfit, there were no differences across distribution conditions. The distribution differences became the most salient when there was moderate degree of misfit, with the highest degree of nonnormality condition yielding the least inflated scaled χ^2. The results highlighted the importance of considering the effect of a design factor in the context of another factor, and no single rule can be thoroughly relied on to evaluate the performance of a model fit index.

Table 5 *ANOVA Results for Scaled* χ^2

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	10016424.03	2	5008212	6201.653	<.001	0.373
N	2756036.22	2	1378018	1706.395	<.001	0.103
dist	2242950.324	2	1121475	1388.719	<.001	0.084
cov * N	3151226.647	4	787806.7	975.539	<.001	0.117
cov * dist	2609697.201	4	652424.3	807.895	<.001	0.097
N * dist	876946.278	4	219236.6	271.48	<.001	0.033
cov * N * dist	901034.278	8	112629.3	139.468	<.001	0.034

Table 6 *Tukey HSD Multiple Comparisons for the Effect of Degree of Model Misfit on Scaled* χ^2

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	-24.064	0.949	<.001	-26.289	-21.838
	c	-102.516	0.951	<.001	-104.746	-100.285
b	c	-78.452	0.952	<.001	-80.683	-76.220

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

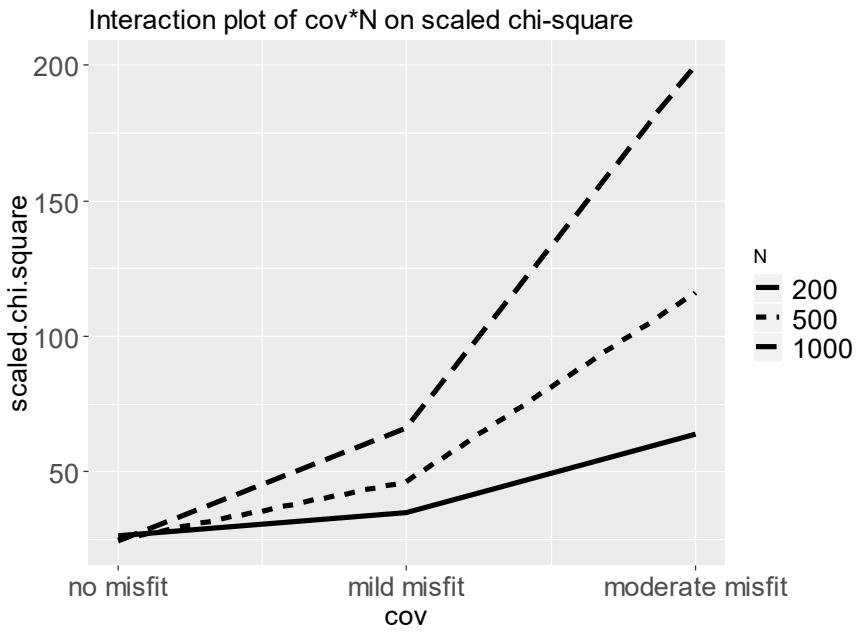
Table 7 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on Scaled* χ^2

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	-20.620	0.953	<.001	-22.853	-18.386
	1000	-55.024	0.953	<.001	-57.257	-52.790
500	1000	-34.404	0.947	<.001	-36.625	-32.183

Table 8 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Scaled* χ^2

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	7.443	0.948	<.001	-26.289	-21.838
	3	46.715	0.952	<.001	-104.746	-100.285
2	3	39.273	0.952	<.001	-80.683	-76.220

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.



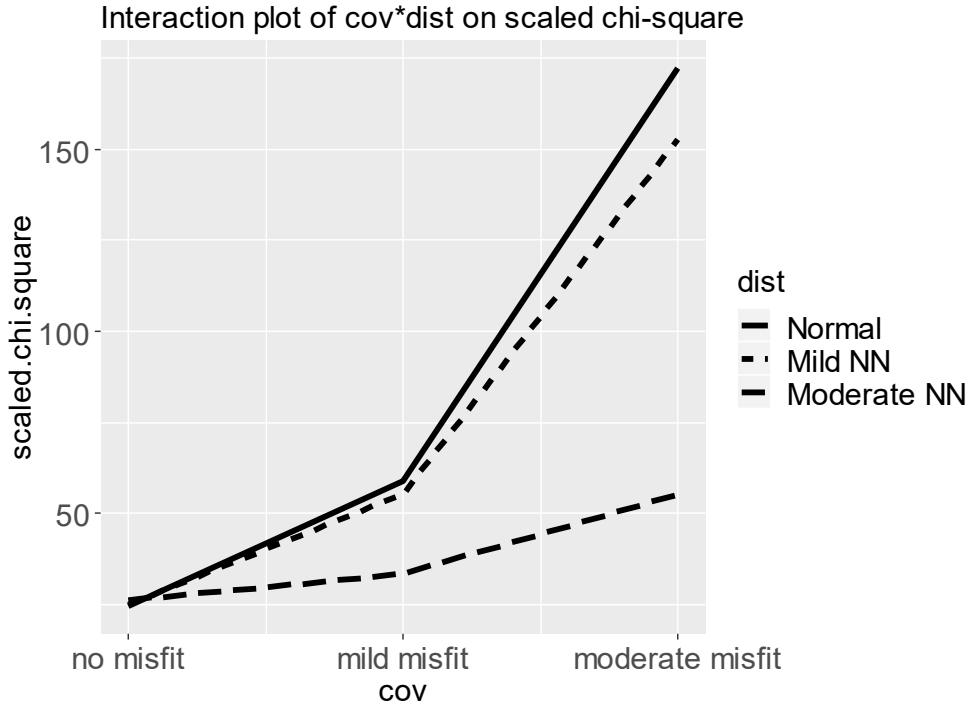


Figure 2. Interaction Plots of Misfit*Sample Size (Upper Panel) and Misfit*Distribution (Lower Panel) for Scaled χ^2 . *Note*. NN: nonnormal.

Table 9 *ANOVA Results for Scaled CFI*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	8.971	2	4.485	3319.02	<.001	0.529
N	0.253	2	0.126	93.602	<.001	0.015
dist	0.006	2	0.003	2.215	0.109	0
cov * N	0.003	4	0.001	0.56	0.692	0
cov * dist	0.258	4	0.064	47.639	<.001	0.015
N * dist	0.24	4	0.06	44.337	<.001	0.014
cov * N * dist	0.013	8	0.002	1.201	0.294	0.001

The ANOVA results for scaled CFI are presented in Table 9. Model misfit was the only design factor that exerted noteworthy effect ($\eta^2 = 0.529$), explaining 52.9% of the variation in the outcome. Tukey's HSD multiple comparison results, as shown in Table 10, revealed that scaled CFI became smaller with increasing level of model misfit, indicating that model misfit impairs model fit, which is within expectation. In the current model, no salient interaction effect was detected. The results indicate that scaled CFI serves as an ideal model fit index given that it is sensitive to model misfit while not saliently affected by other factors such as sample size and distribution of observed variables.

Table 10 *Tukey HSD Multiple Comparisons for the Effect of Degree of Misfit on Scaled CFI*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	0.022	0.001	<.001	-26.289	-21.838
	c	0.096	0.001	<.001	-104.746	-100.285
b	c	0.074	0.001	<.001	-80.683	-76.22

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Similar to scaled CFI, the performance of scaled TLI suggests a decent model index. As shown in Table 11, model misfit was the only predictor that had meaningful practical significance ($\eta^2 = 0.493$) explaining nearly half of the total variation in the outcome. Post hoc multiple comparison results, as shown in Table 12, revealed that scaled TLI became lower with increasing level of model misfit, indicating that the model fit was worse in the presence of misfit. Meanwhile, scaled TLI was robust to other design factors as well as the interactions as indicated by the tiny amount of variation explained by the remaining predictors in the ANOVA model. The results endorsed scaled TLI as another ideal model fit index.

Table 11 *ANOVA Results for Scaled TLI*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	21.999	2	11.000	2757.107	<.001	0.493
N	0.351	2	0.176	44.006	<.001	0.008
dist	0.032	2	0.016	3.962	0.019	0.001
cov * N	0.014	4	0.003	0.852	0.492	0.000
cov * dist	0.459	4	0.115	28.783	<.001	0.010
N * dist	0.399	4	0.100	25.032	<.001	0.009
cov * N * dist	0.054	8	0.007	1.686	0.096	0.001

Table 12 *Tukey HSD Multiple Comparisons for the Effect of Degree of Model Misfit on Scaled TLI*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	0.038	0.002	<.001	0.033	0.043
	c	0.151	0.002	<.001	0.146	0.156
b	c	0.113	0.002	<.001	0.108	0.118

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Table 13 displays ANOVA results for scaled RMSEA. Model misfit accounted for the majority of variation in the outcome ($\eta^2 = 0.604$). Tukey's HSD multiple comparison results, as reported in Table 14, show that scaled RMSEA became larger with increasing level of misfit, which is consistent with theoretical expectation. The finding indicates that the scaled RMSEA is sensitive to misfit, which is a desirable feature. Distribution was another design factor that had practical significance on the outcome ($\eta^2 = 0.096$), explaining about 10% of the variation in the outcome. A further inspection of post hoc multiple comparison (see Table 15) reveals that as the degree of nonnormality increased, scaled RMSEA became smaller, which suggests that scaled RMSEA is sensitive to data nonnormality, especially, in an undesirable direction, posing caution on its performance in evaluating models under nonnormality condition. Above and beyond the main effects, it is more important to discuss the effect of distribution under each level of model misfit, as evidenced by the significant interaction effect between misfit and distribution ($\eta^2 = 0.087$), accounting for almost ten percent of the variation in the outcome. The interaction plot is shown in Figure 3.

Figure 3 reveals that when model is specified correctly, scaled RMSEA was around 0.020 regardless of distribution conditions. When mild degree of misfit was introduced, scaled RMSEA increased to around 0.050 for normal and mild nonnormal distribution conditions, but was around only 0.030 for moderate nonnormal condition. The distribution differences were more salient under moderate model misfit, with the highest degree of nonnormality yielding scaled RMSEA far below the other two distribution conditions. The results suggest that while scaled RMSEA is sensitive to model misfit, its sensitivity is impaired by data nonnormality. The results further demonstrate that the performance of scaled RMSEA in evaluating models should be considered in the context of combined effects of model misfit and data distribution.

Table 13 *ANOVA Results for Scaled RMSEA*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	4.388	2	2.194	8047.260	<.001	0.604
N	0.050	2	0.025	92.254	<.001	0.007
dist	0.697	2	0.349	1279.101	<.001	0.096
cov * N	0.022	4	0.006	20.617	<.001	0.003
cov * dist	0.633	4	0.158	580.513	<.001	0.087
N * dist	0.014	4	0.004	13.035	<.001	0.002
cov * N * dist	0.002	8	0.000	1.047	0.398	0.000

Table 14 *Tukey HSD Multiple Comparisons for the Effect of Degree of Misfit on Scaled RMSEA*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	-0.027	0.001	<.001	-0.029	-0.026
	c	-0.070	0.001	<.001	-0.071	-0.069
b	c	-0.043	0.001	<.001	-0.044	-0.041

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Table 15 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Scaled RMSEA*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	0.003	0.001	<.001	0.001	0.004
	3	0.026	0.001	<.001	0.024	0.027
2	3	0.023	0.001	<.001	0.022	0.024

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

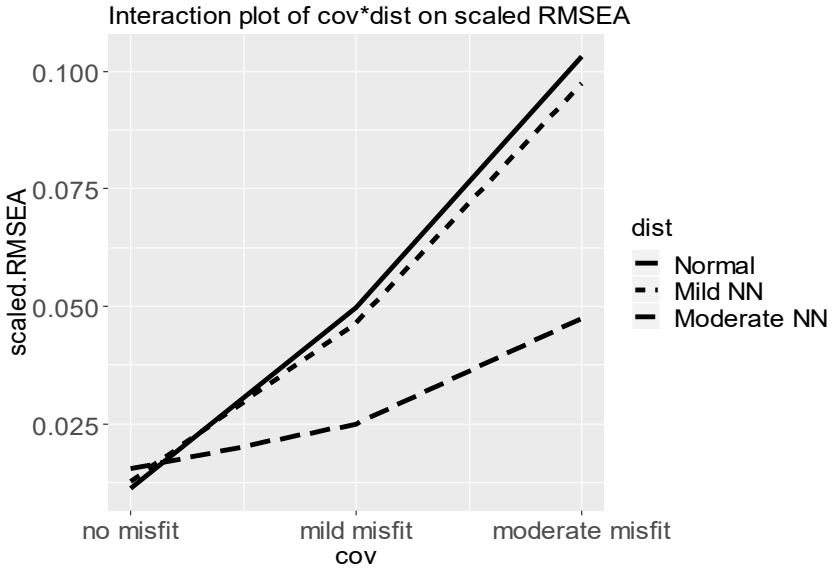


Figure 3. Interaction Plot of Misfit*Distribution for Scaled RMSEA. *Note*. NN: nonnormal.

Table 16 presents the ANOVA results for SRMR. Model misfit and sample size were the noteworthy predictors, accounting for 58.5% and 12.9% of the variation in the outcome respectively. Tukey's HSD multiple comparison results (see Table 17 and Table 18) show that SRMR increased with degree of misfit, and decreased with sample size, indicating that lower misfit and larger sample size contribute to better model fit. More importantly, the interaction between misfit and distribution explained 9.6% of the variation in SRMR. A further inspection of the interaction plot (see Figure 4) reveals that when there was no model misfit, SRMR values were all below .030 across distribution conditions. When misfit was at mild level, SRMR under moderate nonnormal distribution showed the lowest value. The pattern became more salient under moderate misfit condition. The interaction effect indicates that although SRMR is sensitive to model misfit which is desirable, the sensitivity is impaired by a moderate degree of data nonnormality. The results suggest that the performance of SRMR in evaluating models is affected by sample size and data distribution.

Table 16 *ANOVA Results for SRMR*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.976	2	0.488	14009.426	<.001	0.585
N	0.215	2	0.108	3088.169	<.001	0.129
dist	0.091	2	0.045	1299.112	<.001	0.054
cov * N	0.020	4	0.005	139.972	<.001	0.012
cov * dist	0.160	4	0.040	1148.618	<.001	0.096
N * dist	0.009	4	0.002	64.325	<.001	0.005
cov * N * dist	0.001	8	0.000	3.439	0.001	0.001

Table 17 *Tukey HSD Multiple Comparisons for the Effect of Degree of Model Misfit on SRMR*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	-0.012	0.000	<.001	-0.013	-0.012
	c	-0.033	0.000	<.001	-0.033	-0.032
b	c	-0.021	0.000	<.001	-0.021	-0.020

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Table 18 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on SRMR*

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	0.010	0.000	<.001	0.010	0.011
	1000	0.015	0.000	<.001	0.015	0.016
500	1000	0.005	0.000	<.001	0.004	0.005

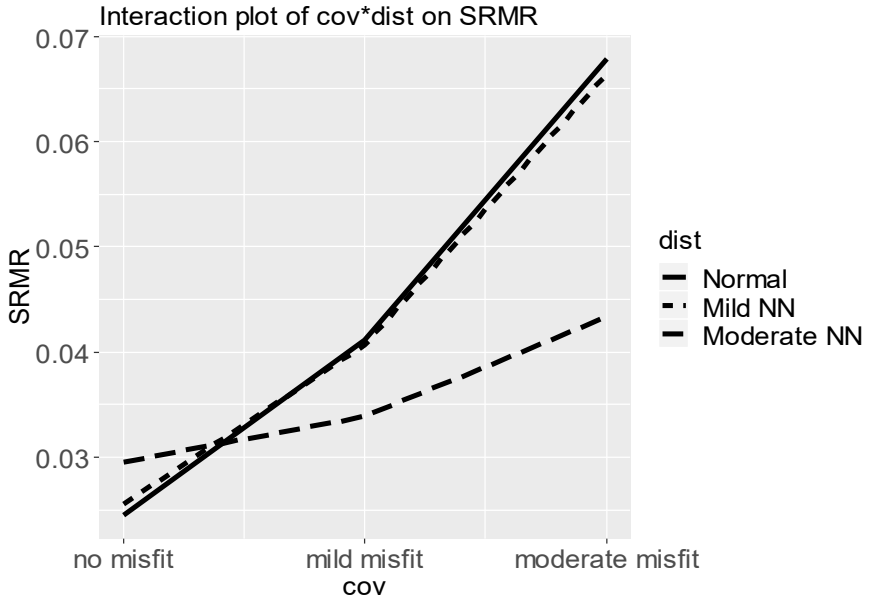


Figure 4. Interaction Plot of Misfit*Distribution for SRMR. *Note*. NN: nonnormal.

Differences between MLR and Bayesian Methods

Rejection Rates of p-values

In an effort to investigate the performance of MLR and Bayesian methods in terms of rejection rates associated with *p* values, descriptive plots were made to check the impacts of design factors on rejection rates. Given that the rejection rate data is aggregated over replicates, ANOVA models are not performed.

Figure 5 through Figure 7 display the rejection rates of *p*-values using MLR and Bayesian estimation methods under the three model misfit conditions. In the basic facet of each figure, x axis refers to degrees of data nonnormality (1, 2, and 3), and y axis refers to rejection rates. The triangular dots denote the rejection rates associated with MLR method, and the circular dots denote the rejection rates associated with Bayesian method. The facets are placed side by side in the increasing order of sample sizes (200, 500, 1000 as shown in the column label). Each figure corresponds to a level of model misfit (ordered as no, mild, moderate misfit).

When the model is correctly specified (see Figure 5), it would be ideal that rejection rates fall below .05 regardless of sample size and data nonnormality, and thus a reference line was flagged at .05 in the figure. Under small sample size ($N = 200$), both MLR and Bayesian estimation methods showed greater rejection rates under data nonnormality. With larger sample size ($N = 500$ or 1000), rejection rates of *p* value (PP *p*-value for Bayesian) using either method were not affected much by distribution conditions. The performance of rejection rates appeared to have less fluctuations under

larger sample sizes. Contrasting the two estimation methods, in general, *p* values associated with MLR had higher rejection rates, most of which being larger than .05. The most notable difference between the two methods was observed when sample size was small, as can be told from the gap between the circular and triangular dots. Especially, rejection rates under MLR estimation exhibited the highest value when sample size was small and degree of nonnormality was moderate. The results suggest that under a correctly specified model, *p* values associated with MLR estimation in general have inflated type I error across conditions of other design factors, and the worst performance occurs under the combined condition of small sample size and nonnormal data distribution. PP *p*-values associated with Bayesian estimation have stable performance across conditions of design factors, and yield rejection rates consistently below .050. The finding lends support for Bayesian method in terms of decent type I error control compared to MLR method, especially when sample size is small.

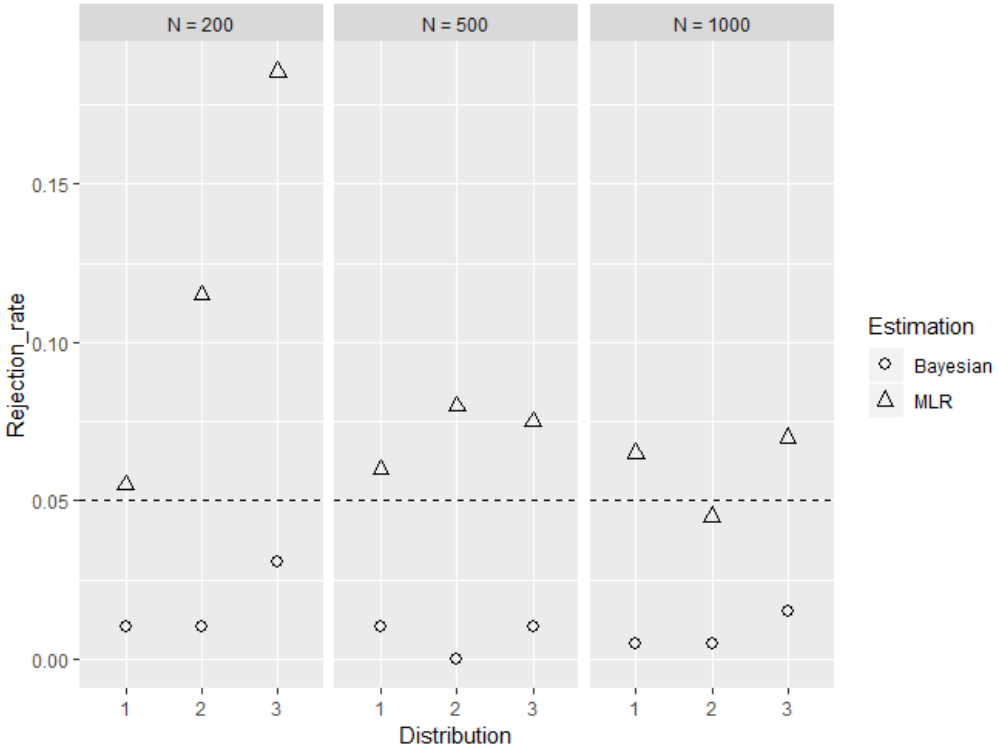


Figure 5. Rejection Rates of *p*-values Associated with MLR and Bayesian under No Misfit. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

When the model is mildly misspecified (see Figure 6), it is expected that rejection rates be higher than those in Figure 5 but still below 80% because mild degree of misspecification is acceptable and realistic, so a reference line was flagged at .80 in the figure. For both methods, rejection rates increased with sample size and became smaller when data nonnormality reached moderate level. In general, Bayesian method produced lower rejection rates than MLR. Under small sample size ($N = 200$), all the rejection rates were below .80. When sample sizes were larger and data were not deviating much from normality, rejection rates for both methods exceeded 80%. Comparing Bayesian and MLR methods, the most notable difference was observed when sample size was small. The results suggest that under mildly misspecified models, rejection rates of both methods are positively related to sample size and negatively related to data nonnormality (moderate vs less degrees of nonnormality). MLR associated *p* values over-reject the model under larger sample sizes, and Bayesian associated PP *p* values exhibit decent performance when sample size is smaller than 1000.

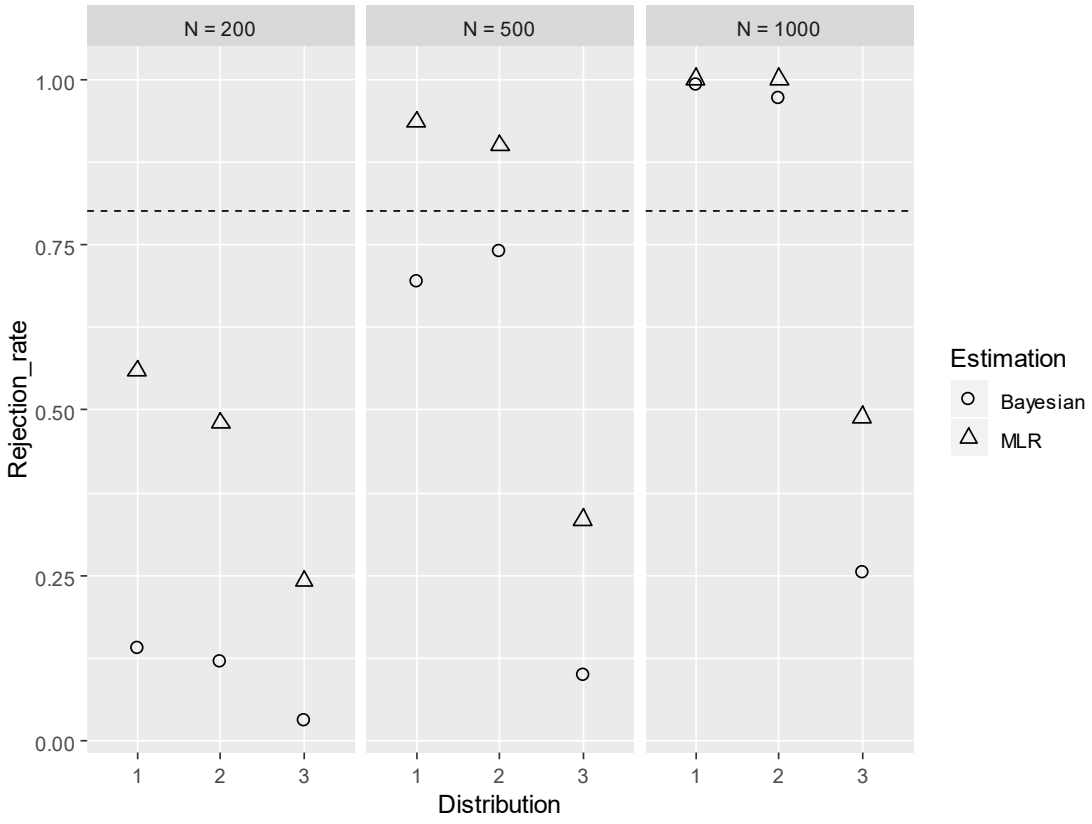


Figure 6. Rejection Rates of *p*-values Associated with MLR and Bayesian Method under Mild Misfit. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

When the model is moderately misspecified (see Figure 7), it would be ideal that rejection rates fall above 80% because the model is expected to be rejected, so a reference line was flagged at .80 in the figure. For both methods, rejection rates exceeded 80% regardless of sample sizes when data distribution were not deviating from normality much. When data were moderately misspecified and sample sizes were smaller than 1000, rejection rates were deflated, indicating an impaired statistical power to detect the misfit. Comparing the two methods, they generally yielded similar results. The findings suggest that when the model is moderately misspecified, both Bayesian and MLR methods exhibit similar and decent statistical power to reject the model even under small sample sizes ($N = 200$), except for the conditions of moderate degree of data nonnormality.

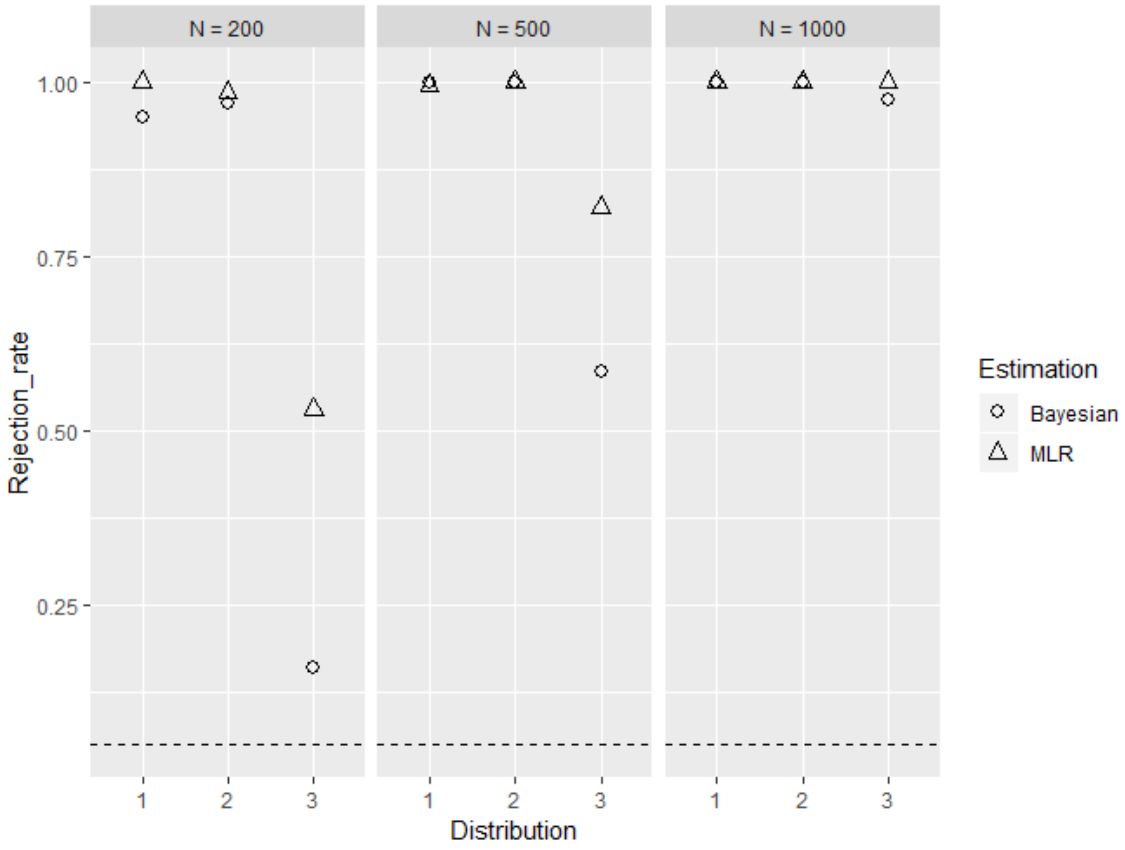


Figure 7. Rejection Rates of *p*-values Associated with MLR and Bayesian under Moderate Misfit. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Relative Bias in Point Estimates

In order to investigate the performance of MLR and Bayesian methods in terms of point estimations, descriptive plots were made to check the impacts of design factors on each outcome, and a series of factorial ANOVA models were performed to examine the effects of the design factors and their interactions on relative bias of loadings and inter-factor correlations associated with MLR and Bayesian method separately. In an initial effort of inspecting the descriptive statistics (mean, standard deviation, minimum, maximum, and distributions) of the parameter estimates (point estimates, and standard error estimates) based on the simulated data, the parameter estimates were comparable across items (for loading related parameters) or latent factors (for inter-factor correlation related parameters), and thus relative biases of estimates were calculated based on an arbitrarily selected parameter associated with the loading from the 2nd item to the 1st latent factor for loading estimates, and the inter-factor correlation between the first two latent factors for covariance estimates. This applies to all the subsequent analyses related to parameter estimations. Practical significance was represented by η^2, and those effects with η^2 larger than or close to .10 were considered as having adequate contribution to the model. Tukey's HSD multiple comparisons were conducted for the noticeable main effects. For the interaction effects with adequate practical significance, interaction plots based on marginal means were further inspected to gain a better understanding of the relationships.

Figure 8 presents the relative biases of loading estimates using MLR and Bayesian estimation methods across the design factor conditions. Each row represents a

level of model misfit (in increasing order), and each column represents a level of sample size (ordered from 200 to 1000). In each basic facet, x axis refers to increasing degree of data nonnormality (denoted as dist), and y axis refers to relative biases. A reference line was flagged at the value of .10 to indicate considerable magnitude of relative bias. Circular dots denote Bayesian estimation and triangular dots denote MLR estimation. In general, relative biases became slightly smaller as sample size increased. Compared to normal and mildly nonnormal data distributions, moderately nonnormal distribution produced higher relative biases, indicating that data nonnormality is associated with parameter inaccuracy. When distributions were normal or mildly nonnormal, relative biases were below or close to .10 across sample sizes and estimation methods, indicating an acceptable level of parameter accuracy. Comparing the three misfit conditions, models with higher degree of misfit appeared to produce slightly lower parameter accuracy. Comparing MLR and Bayesian estimation methods, parameter accuracy did not seem to be affected by the two methods. The most notable (if any) differences occurred under small sample size ($N = 200$) combined with moderate nonnormality. The results indicate that under normal or mildly nonnormal data distributions, both MLR and Bayesian estimation methods produce acceptable parameter accuracy for loading estimations.

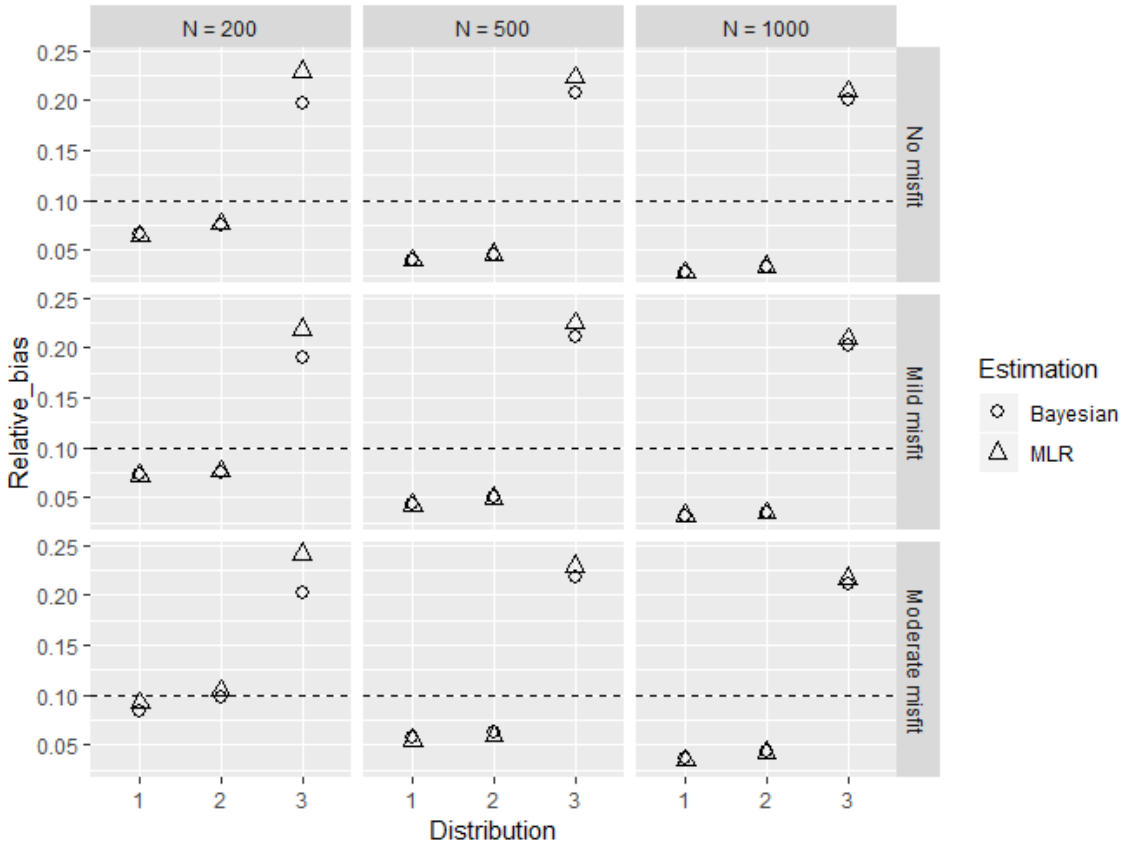


Figure 8. Relative Biases of Loadings Associated with MLR and Bayesian across Conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 19 shows the ANOVA results for relative bias of loadings estimated using MLR. It can be seen that data distribution was the only noteworthy predictor, accounting for half of the variation in the outcome (η^2 = .500). The remaining main effects and interaction effects did not exert remarkable contribution. A further inspection of Tukey's HSD multiple comparison results (see Table 20) reveal that higher degree of data nonnormality was associated with greater relative bias in loadings. The result indicates

that relative bias of loadings using MLR estimation method is mainly affected by data distribution, and data nonnormality is related to lower parameter accuracy.

Table 19 *ANOVA Results for Relative Bias of Loadings with MLR Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.213	2	0.106	17.923	<.001	0.003
N	1.253	2	0.626	105.596	<.001	0.019
dist	33.397	2	16.699	2815.329	<.001	0.500
cov * N	0.057	4	0.014	2.413	0.047	0.001
cov * dist	0.015	4	0.004	0.646	0.630	0.000
N * dist	0.210	4	0.053	8.858	<.001	0.003
cov * N * dist	0.015	8	0.002	0.321	0.959	0.000
Total	66.798	5360				

Table 20 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Relative Bias of Loadings with MLR Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.007	0.003	<.001	-0.013	-0.001
	3	-0.171	0.003	<.001	-0.177	-0.165
2	3	-0.164	0.003	<.001	-0.170	-0.158

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 21 shows the ANOVA results for relative bias of loadings estimated using Bayesian method. It can be seen that data distribution was the only noteworthy predictor, accounting for nearly half of the variation in the outcome (η^2 = .459). The remaining

main effects and interaction effects did not have notable contribution. A further inspection of Tukey's HSD multiple comparison results (see Table 22) reveal that higher degree of data nonnormality was associated with greater parameter inaccuracy. The result indicates that relative bias of loadings using Bayesian estimation method is mainly affected by data nonnormality. Comparing relative bias of loadings with MLR and Bayesian estimation methods, Bayesian method yielded slightly lower overall variation in relative bias (Type III SS = 58.070) than that of MLR method (Type III SS = 66.798). Data nonnormality was the only noteworthy predictor for both ANOVA models, and it accounted for slightly less proportion of variation in the outcome estimated with Bayesian method (η^2 = .459) than with MLR method (η^2 = .500). The results from Table 19 through Table 22 suggest that MLR and Bayesian estimation methods have similar performance in terms of parameter accuracy in loadings.

Table 21 *ANOVA Results for Relative Bias of Loadings with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.183	2	0.092	16.287	<.001	0.003
N	0.628	2	0.314	55.826	<.001	0.011
dist	26.642	2	13.321	2368.293	<.001	0.459
cov * N	0.009	4	0.002	0.389	0.817	0.000
cov * dist	0.013	4	0.003	0.586	0.673	0.000
N * dist	0.578	4	0.145	25.711	<.001	0.010
cov * N * dist	0.013	8	0.002	0.297	0.967	0.000
Total	58.070	5360				

Table 22 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Relative Bias of Loadings with Bayesian Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.007	0.003	0.025	-0.012	-0.001
	3	-0.153	0.003	<.001	-0.159	-0.147
2	3	-0.147	0.003	<.001	-0.153	-0.141

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Figure 9 presents the relative biases of inter-factor correlations using MLR and Bayesian estimation methods across conditions. The figure was displayed in the same format as Figure 8. A reference line was flagged at the value of .10 to indicate considerable magnitude of relative bias. Circular dots refer to Bayesian estimation and triangular dots refer to RML estimation. In general, relative biases decreased slightly as sample size increased. Compared to normal and mildly nonnormal data distributions, moderately nonnormal distribution conditions were associated with higher parameter inaccuracy. An acceptable level of parameter accuracy was only observed when sample sizes reached 1000, where relative biases were below or close to .10. Comparing the three misfit conditions, relative biases were slightly affected. Comparing MLR and Bayesian estimation methods, parameter accuracy did not seem to be affected by the two methods given that the circular and triangular dots were generally overlapping. The results indicate that parameter accuracy of inter-factor correlations is impaired by moderate data nonnormality, and an acceptable level of relative biases can only be reached under large sample sizes.

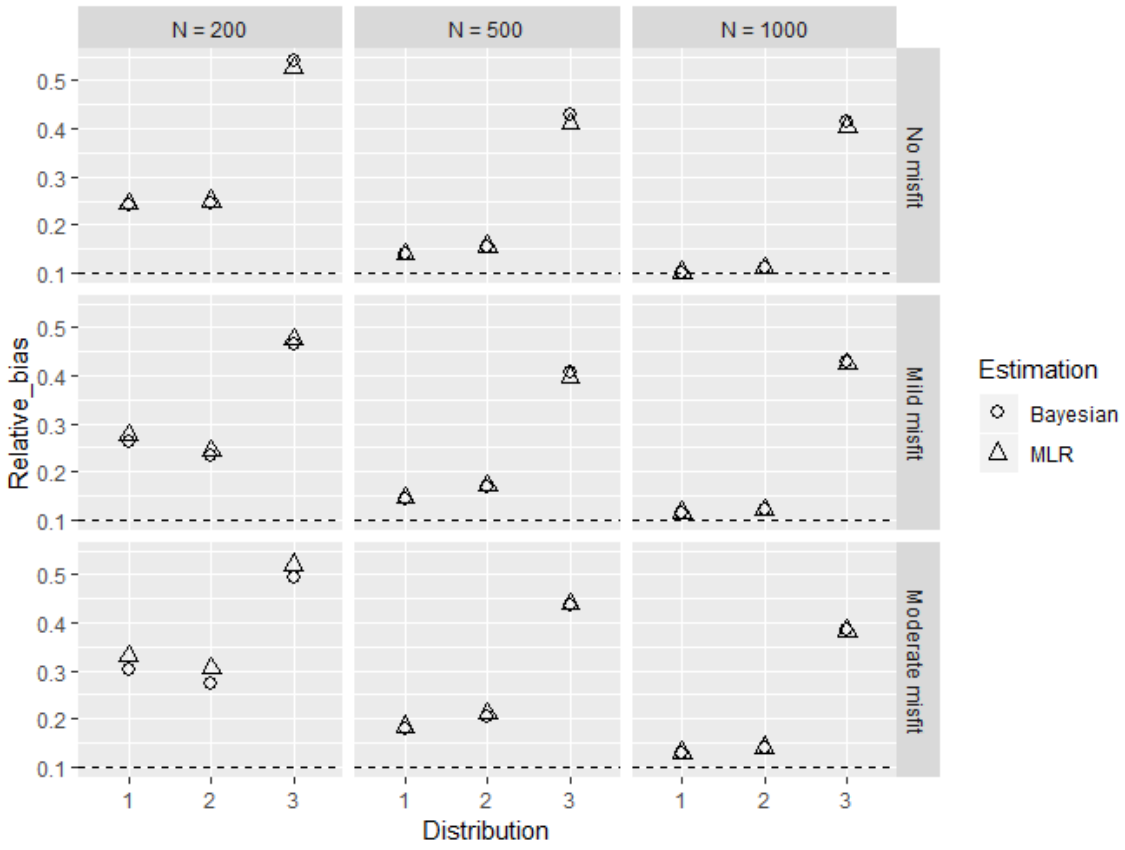


Figure 9. Relative Biases of Inter-Factor Correlations Associated with MLR and Bayesian across Conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 23 shows the ANOVA results for relative bias of inter-factor correlations estimated using MLR. It can be seen that data distribution was the most noteworthy predictor, accounting for 24.6% of the variation in the outcome (η^2 = .246). Sample size accounted for a small amount of the overall variation (η^2 = .058). The remaining main effects and interaction effects only exerted little contribution. A further inspection of Tukey's HSD multiple comparison results (see Table 24) reveal that moderate degree of

data nonnormality was associated with greater relative bias in inter-factor correlations than the other two distribution conditions. The results indicate that parameter accuracy of inter-factor correlations using MLR estimation method is mainly affected by data distribution, and moderate data nonnormality would impair parameter accuracy of inter-factor correlations.

Table 23 *ANOVA Results for Relative Bias of Inter-Factor Correlations with MLR Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	1.206	2	0.603	14.969	<.001	0.004
N	18.199	2	9.099	225.923	<.001	0.058
dist	76.962	2	38.481	955.430	<.001	0.246
cov * N	0.520	4	0.130	3.226	0.012	0.002
cov * dist	0.495	4	0.124	3.071	0.015	0.002
N * dist	0.802	4	0.201	4.979	0.001	0.003
cov * N * dist	0.358	8	0.045	1.110	0.353	0.001
Total	313.373	5360				

Table 24 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Relative Bias of Inter-Factor Correlations with MLR Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.004	0.007	0.813	-0.020	0.012
	3	-0.256	0.007	<.001	-0.272	-0.240
2	3	-0.252	0.007	<.001	-0.268	-0.236

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 25 shows the ANOVA results for relative bias of inter-factor correlations estimated using Bayesian method. It can be seen that data distribution was the only noteworthy predictor, accounting for 27.6% of the variation in the outcome (η^2 = .276). Sample size accounted for a small amount of the overall variation (η^2 = .048). The remaining main effects and interaction effects exerted little contribution. A further inspection of Tukey's HSD multiple comparison results (see Table 26) reveal that moderate degree of data nonnormality was associated with greater parameter inaccuracy. The result indicates that relative bias of inter-factor correlations using Bayesian estimation method is mainly affected by data nonnormality. Comparing MLR and Bayesian estimation methods, Bayesian method yielded similar overall variation in relative bias (Type III SS = 298.335) than that of MLR method (Type III SS = 313.373). Data nonnormality was the only noteworthy predictor for both ANOVA models, and it accounted for comparable proportion of variation in the outcome estimated with Bayesian method (η^2 = .276) than with MLR method (η^2 = .246). The results from Table 23 through Table 26 suggest that MLR and Bayesian estimation methods have comparable performance in terms of parameter accuracy in inter-factor correlations.

Table 25 *ANOVA Results for Relative Bias of Inter-Factor Correlations with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.478	2	0.239	6.410	0.002	0.002
N	14.440	2	7.220	193.824	<.001	0.048
dist	82.399	2	41.199	1105.987	<.001	0.276
cov * N	0.369	4	0.092	2.477	0.042	0.001
cov * dist	0.846	4	0.212	5.681	<.001	0.003
N * dist	0.729	4	0.182	4.894	0.001	0.002
cov * N * dist	0.376	8	0.047	1.262	0.259	0.001
Total	298.335	5360				

Table 26 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Relative Bias of Inter-Factor Correlations with Bayesian Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.004	0.006	0.841	-0.019	0.011
	3	-0.265	0.006	<.001	-0.280	-0.250
2	3	-0.261	0.006	<.001	-0.276	-0.246

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Standard Errors

In order to examine the performance of MLR and Bayesian in terms of standard errors, descriptive plots were made to visualize the impacts of design factors on the standard errors of loading and inter-factor correlations, and factorial ANOVA models were fit separately for the related outcomes.

Figure 10 presents the standard errors of loadings using MLR and Bayesian estimation methods across conditions. The figure is displayed in the same format as Figure 8. Circular and triangular dots refer to Bayesian estimation and MLR estimation respectively. In general, standard errors decreased as sample size increased. Moderately nonnormal distribution conditions produced larger standard errors compared to the less nonnormal distribution conditions. Comparing the three misfit conditions, no salient differences in relative biases were detected. Comparing MLR and Bayesian estimation methods, the circular dots overlapped with the triangular dots in some conditions, and fell slightly below the triangular dots in some other conditions. Explicitly, the most remarkable gap between the two methods occurred when sample size was 200, data distributions were moderately nonnormal, and models were moderately misspecified. The results indicate that standard errors of loadings are mainly affected by sample size and data nonnormality, and Bayesian method produces smaller standard errors than MLR method in the combined condition of small sample size, higher degrees of nonnormality and model misfit.

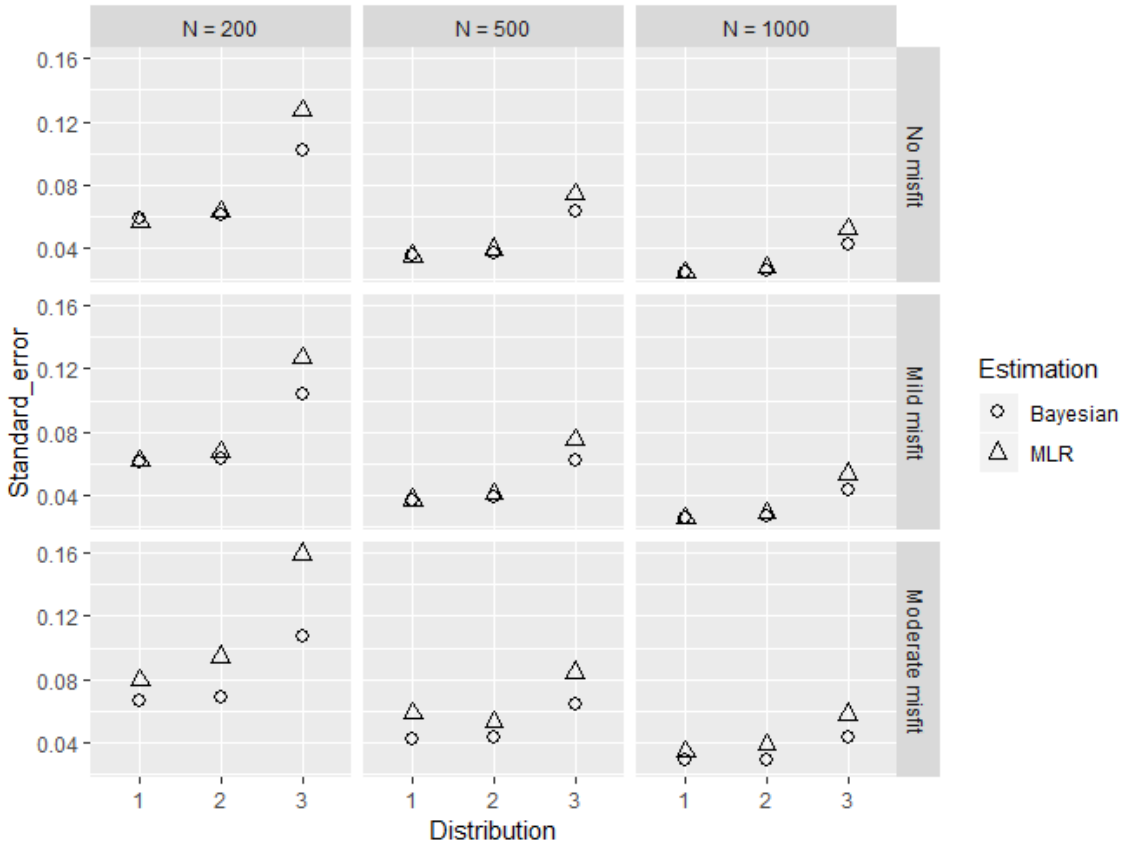


Figure 10. Standard Errors of Loadings Associated with MLR and Bayesian across Conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 27 presents the ANOVA results for standard errors of loadings estimated using MLR. It can be seen that sample size and data distribution were the most noteworthy predictors, accounting for 15.7% and 11.9% of the variation in the outcome respectively (η^2 = .157 for *N*, and η^2 =.119 for dist). The remaining main effects and interaction effects only exerted a small contribution. A further inspection of Tukey's HSD multiple comparison results (see Tables 28 and 29) show that larger sample sizes were associated with smaller standard errors of loadings, and higher degree of data

nonnormality was associated with larger standard errors. The results indicate that standard errors of loadings using MLR estimation method are mainly affected by sample size and data distribution, and larger sample sizes and greater approximation to normal distribution contribute to smaller standard errors of loadings.

Table 27 *ANOVA Results for Standard Errors of Loadings with MLR Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.335	2	0.168	74.590	<.001	0.019
N	2.783	2	1.391	619.114	<.001	0.157
dist	2.099	2	1.050	467.096	<.001	0.119
cov * N	0.075	4	0.019	8.314	<.001	0.004
cov * dist	0.001	4	0.000	0.137	0.969	0.000
N * dist	0.404	4	0.101	44.957	<.001	0.023
cov * N * dist	0.024	8	0.003	1.333	0.222	0.001
Total	17.708	5360				

Table 28 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on Standard Errors of Loadings with MLR Estimator*

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	0.037	0.002	<.001	0.033	0.040
	1000	0.054	0.002	<.001	0.050	0.057
500	1000	0.017	0.002	<.001	0.013	0.021

Table 29 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Standard Errors of Loadings with MLR Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.005	0.002	0.011	-0.008	-0.001
	3	-0.043	0.002	<.001	-0.047	-0.039
2	3	-0.039	0.002	<.001	-0.042	-0.035

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 30 presents the ANOVA results for standard errors of loadings estimated using Bayesian method. It can be seen that sample size accounted for a substantial proportion of overall variation (η^2 = .550), followed by data distribution which accounted for 26.0% of the variation (η^2 =.260). The remaining main effects and interaction effects only exerted small contribution. A further inspection of Tukey's HSD multiple comparison results (see Tables 31 and 32) reveal that larger sample sizes and less deviation from normal data distribution were associated with smaller standard errors. Comparing MLR and Bayesian estimation methods, Bayesian method yielded considerably smaller overall variation in standard errors (Type III SS = 3.361) than that of MLR method (Type III SS = 17.708). Sample size and data nonnormality, the noteworthy main effects in both ANOVA models, accounted for considerably higher proportion of variation in the outcome estimated with Bayesian method than with MLR method. The results suggest that standard errors of loadings estimated by MLR and Bayesian methods are both affected by sample size and data distribution, and Bayesian method yields less variation in standard errors.

Table 30 *ANOVA Results for Standard Errors of Loadings with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.020	2	0.010	111.452	<.001	0.006
N	1.850	2	0.925	10078.793	<.001	0.550
dist	0.873	2	0.436	4753.927	<.001	0.260
cov * N	0.003	4	0.001	7.323	<.001	0.001
cov * dist	0.003	4	0.001	7.579	<.001	0.001
N * dist	0.122	4	0.030	332.212	<.001	0.036
cov * N * dist	0.001	8	0.000	0.877	0.535	0.000
Total	3.361	5360				

Table 31 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on Standard Errors of Loadings with Bayesian Estimator*

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	0.030	0.0003	<.001	0.029	0.030
	1000	0.044	0.0003	<.001	0.043	0.045
500	1000	0.015	0.0003	<.001	0.014	0.015

Table 32 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Standard Errors of Loadings with Bayesian Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.001	0.000	<.001	-0.002	-0.0007
	3	-0.027	0.000	<.001	-0.028	-0.026
2	3	-0.026	0.000	<.001	-0.027	-0.025

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Figure 11 presents the standard errors of inter-factor correlations using MLR and Bayesian estimation methods across conditions. Circular and triangular dots denote Bayesian and MLR estimation method respectively. In general, standard errors decreased as sample size increased. Similar to the standard errors for loadings, moderately nonnormal distribution conditions yielded larger standard errors compared to the distribution conditions deviating less from normality. Comparing the three misfit conditions, the relative biases were slightly affected. Comparing MLR and Bayesian estimation methods, the circular dots generally overlapped with the triangular dots under correctly or mildly misspecified models, and fell slightly below the triangular dots under moderately misspecified models. Similar to the pattern in Figure 10, the greatest gap between the two methods occurred in the combined condition of small sample size, moderate data nonnormality and moderate model misfit. The results indicate that standard errors of inter-factor correlations are mainly affected by sample size and data distribution, and Bayesian method produces smaller standard errors than MLR method under moderate model misfit conditions.

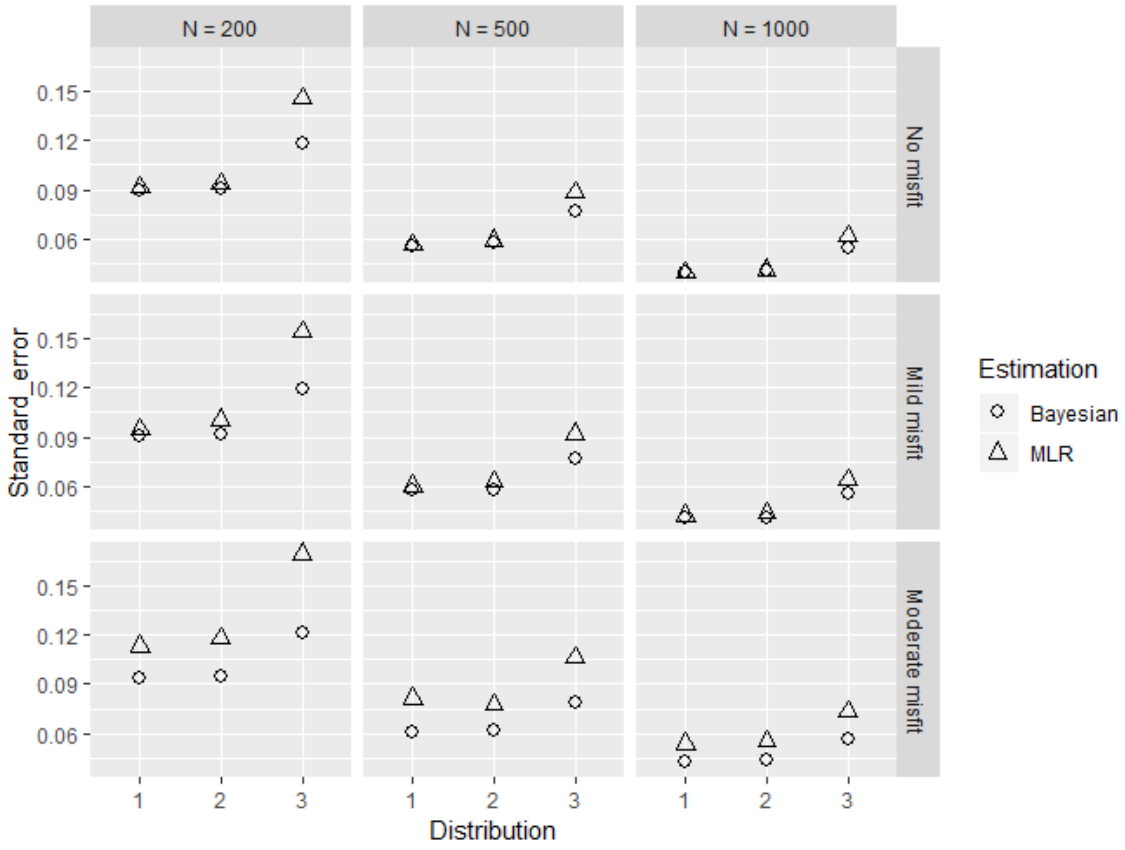


Figure 11. Standard errors of inter-factor correlations associated with MLR and Bayesian across conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 33 presents the ANOVA results for standard errors of inter-factor correlations estimated using MLR. Similar to the ANOVA results for standard errors of loadings, sample size and data distribution were the most noteworthy predictors, accounting for 41.8% and 14.4% of the variation in the outcome (η^2 = .418 for *N*, and η^2 =.144 for dist). The remaining main effects and interaction effects only made a small contribution. A further inspection of Tukey's HSD multiple comparison results (see

Tables 34 and 35) show that larger sample sizes and less deviation from normal distribution were associated with smaller standard errors.

Table 33 *ANOVA Results for Standard Errors of Inter-Factor Correlations with MLR Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.355	2	0.177	256.708	<.001	0.036
N	4.112	2	2.056	2975.799	<.001	0.418
dist	1.419	2	0.709	1026.677	<.001	0.144
cov * N	0.016	4	0.004	5.694	<.001	0.002
cov * dist	0.002	4	0.001	0.806	0.521	0.000
N * dist	0.247	4	0.062	89.332	<.001	0.025
cov * N * dist	0.004	8	0.000	0.646	0.739	0.000
Total	9.840	5360				

Table 34 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on Standard Errors of Inter-Factor Correlation with MLR Estimator*

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	0.043	0.0009	<.001	0.041	0.045
	1000	0.066	0.0009	<.001	0.064	0.068
500	1000	0.023	0.0009	<.001	0.021	0.025

Table 35 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Standard Errors of Inter-Factor Correlation with MLR Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.002	0.001	0.048	-0.004	-0.00001
	3	-0.035	0.001	<.001	-0.037	-0.033
2	3	-0.033	0.001	<.001	-0.035	-0.030

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 36 presents the ANOVA results for standard errors of inter-factor correlations estimated using Bayesian method. It can be seen that sample size accounted for a substantial proportion of overall variation (η^2 = .795), followed by data distribution which explained 14.3% of the variation (η^2 =.143). The remaining main effects and interaction effects exerted little contribution. Tukey's HSD multiple comparison results (see Table 37 and Table 38) reveal that standard errors decreased as sample size increased and as data approached a normal distribution. Comparing MLR and Bayesian estimation methods, Bayesian method yielded smaller overall variation in standard errors (Type III SS = 3.456) than that of MLR method (Type III SS = 9.840). Sample size and data nonnormality exerted noteworthy main effects in both ANOVA models. Sample size accounted for considerably higher proportion of variation in the outcome estimated with Bayesian method than with MLR method. The results suggest that standard errors of inter-factor correlations estimated by MLR and Bayesian methods are both affected by sample size and data distribution, and Bayesian method produced less variation in standard errors.

Table 36 *ANOVA Results for Standard Errors of Inter-Factor Correlations with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	0.011	2	0.005	174.137	<.001	0.003
N	2.747	2	1.374	43625.435	<.001	0.795
dist	0.494	2	0.247	7845.655	<.001	0.143
cov * N	0.000	4	0.000	2.302	0.056	0.000
cov * dist	0.000	4	0.000	3.800	0.004	0.000
N * dist	0.035	4	0.009	278.340	<.001	0.010
cov * N * dist	0.000	8	0.000	0.834	0.572	0.000
Total	3.456	5360				

Table 37 *Tukey HSD Multiple Comparisons for the Effect of Sample Size on Inter-Factor Correlation with Bayesian Estimator*

(I) N	(J) N	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
200	500	0.036	0.0002	<.001	0.035	0.036
	1000	0.054	0.0002	<.001	0.054	0.055
500	1000	0.019	0.0002	<.001	0.018	0.019

Table 38 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Inter-Factor Correlation with Bayesian Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.001	0.000	<.001	-0.001	-0.0005
	3	-0.020	0.000	<.001	-0.021	-0.020
2	3	-0.019	0.000	<.001	-0.020	-0.019

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Relative Biases of Standard Errors

With the aim of examining the accuracy of standard error estimations using MLR and Bayesian methods, descriptive plots were made to visually inspect the effects of design factors on the relative biases in standard errors, factorial ANOVA models were performed for relative biases in standard errors of loadings and inter-factor correlations separately.

Figure 12 presents the relative biases of standard errors of loadings using MLR and Bayesian estimation methods across conditions. Circular and triangular dots refer to Bayesian and RML estimation respectively. A reference line was flagged at the value of .10 to indicate considerable magnitude of relative bias. Regarding MLR method, it appeared not be affected by the design factors remarkably. When models were correctly specified or mildly misspecified, relative biases yielded by MLR were consistently below .010, indicating a non-severe degree of bias. In contrast, relative biases associated with Bayesian method increased with sample size, and generally became larger with increasing degrees of data nonnormality. Under lower degrees of model misfit (cov = a or b) and smaller sample size conditions (N = 200 or 500), Bayesian method produced relative biases of standard errors less than or close to .10. The results indicate that relative bias of standard errors of loadings associated with Bayesian method were more sensitive to the design factors of model misfit and data distribution compared to those yielded by MLR method.

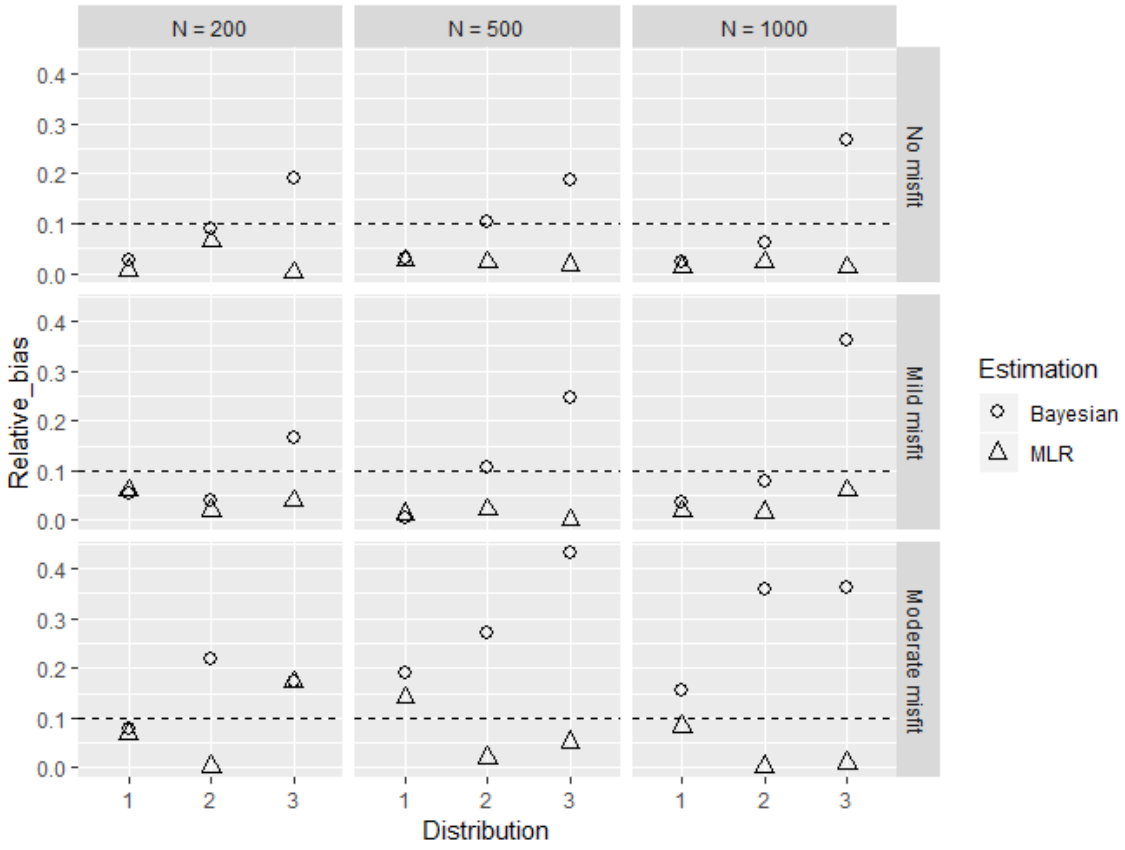


Figure 12. Relative bias of standard errors of loadings associated with MLR and Bayesian across conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 39 presents the ANOVA results for relative bias of standard errors of loadings estimated using MLR method. It can be seen that the main effects and interaction effects exerted little contribution, with the largest η^2 being .033. The corresponding results estimated using Bayesian method are presented in Table 40. Model misfit and data distribution served as the most noteworthy predictors, accounting for 15.4 % and 18.9% of the variation in the outcome (η^2 = .154 and .189 respectively). Tukey's HSD multiple comparison results (see Tables 41 and 42) reveal that relative bias

of standard errors increased as degree of model misfit and data nonnormality increased. Comparing MLR and Bayesian estimation methods, Bayesian method yielded smaller overall variation in relative bias of standard errors (Type III SS = 69.432) than that of MLR method (Type III SS = 1530.135). Results associated with Bayesian method were affected by model misfit and data nonnormality, while results associated with MLR did not appear to be affected by the design factors much. The results suggest that relative bias of standard errors of loadings estimated by MLR and Bayesian methods differ in terms of influencing factors, and Bayesian method produced less variation in relative bias of standard errors.

Table 39 *ANOVA Results for Relative Bias of Standard Errors of Loadings with MLR*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	50.013	2	25.006	91.855	<.001	0.033
N	14.143	2	7.072	25.976	<.001	0.009
dist	0.415	2	0.208	0.762	0.467	0.000
cov * N	1.605	4	0.401	1.474	0.207	0.001
cov * dist	4.591	4	1.148	4.216	0.002	0.003
N * dist	3.598	4	0.900	3.304	0.010	0.002
cov * N * dist	3.661	8	0.458	1.681	0.098	0.002
Total	1530.135					

Table 40 *ANOVA Results for Relative Bias of Standard Errors of Loadings with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	10.686	2	5.343	711.771	<.001	0.154
N	0.068	2	0.034	4.525	0.011	0.001
dist	13.104	2	6.552	872.811	<.001	0.189
cov * N	1.639	4	0.410	54.589	<.001	0.024
cov * dist	1.806	4	0.451	60.137	<.001	0.026
N * dist	1.158	4	0.289	38.559	<.001	0.017
cov * N * dist	0.929	8	0.116	15.466	<.001	0.013
Total	69.432					

Table 41 *Tukey HSD Multiple Comparisons for the Effect of Model Misfit on Relative Bias of Standard Errors of Loadings with Bayesian Estimator*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	-0.010	0.003	0.001	-0.017	-0.003
	c	-0.100	0.003	<.001	-0.107	-0.093
b	c	-0.090	0.003	<.001	-0.097	-0.083

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

Table 42 *Tukey HSD Multiple Comparisons for the Effect of Distribution on Relative Bias of Standard Errors of Loadings with Bayesian Estimator*

(I) dist	(J) dist	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
1	2	-0.041	0.003	<.001	-0.048	-0.035
	3	-0.120	0.003	<.001	-0.127	-0.113
2	3	-0.079	0.003	<.001	-0.086	-0.072

Note. dist: data distribution, 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Figure 13 presents the relative biases of standard errors of inter-factor correlations using the two estimation methods across conditions. Circular and triangular dots denote Bayesian and RML estimation respectively. A reference line was flagged at the value of .10 to indicate considerable relative bias. The patterns were less clear than those in the previous plots, which suggested that the performance of both methods need to be considered under the combined effects of the design factors. Under correctly specified models, the relative biases were generally below .010, and were merely affected by sample sizes and data distributions. Under mildly misspecified models, the relative biases produced by MLR were below or close to .010 across conditions. In contrast, the relative biases produced by Bayesian method decreased with data nonnormality when $N = 200$, and increased with data nonnormality under larger sample sizes. Under moderately specified models, MLR method produced most ideal results when $N = 1000$, while Bayesian method produced best performance when $N = 200$ and data were nonnormally distributed. The results indicate that in terms of accuracy of standard errors for inter-factor correlations, Bayesian method is preferred when sample size is small and data are nonnormally distributed, while MLR method is preferred under large sample sizes ($N = 1000$) regardless of data distribution.

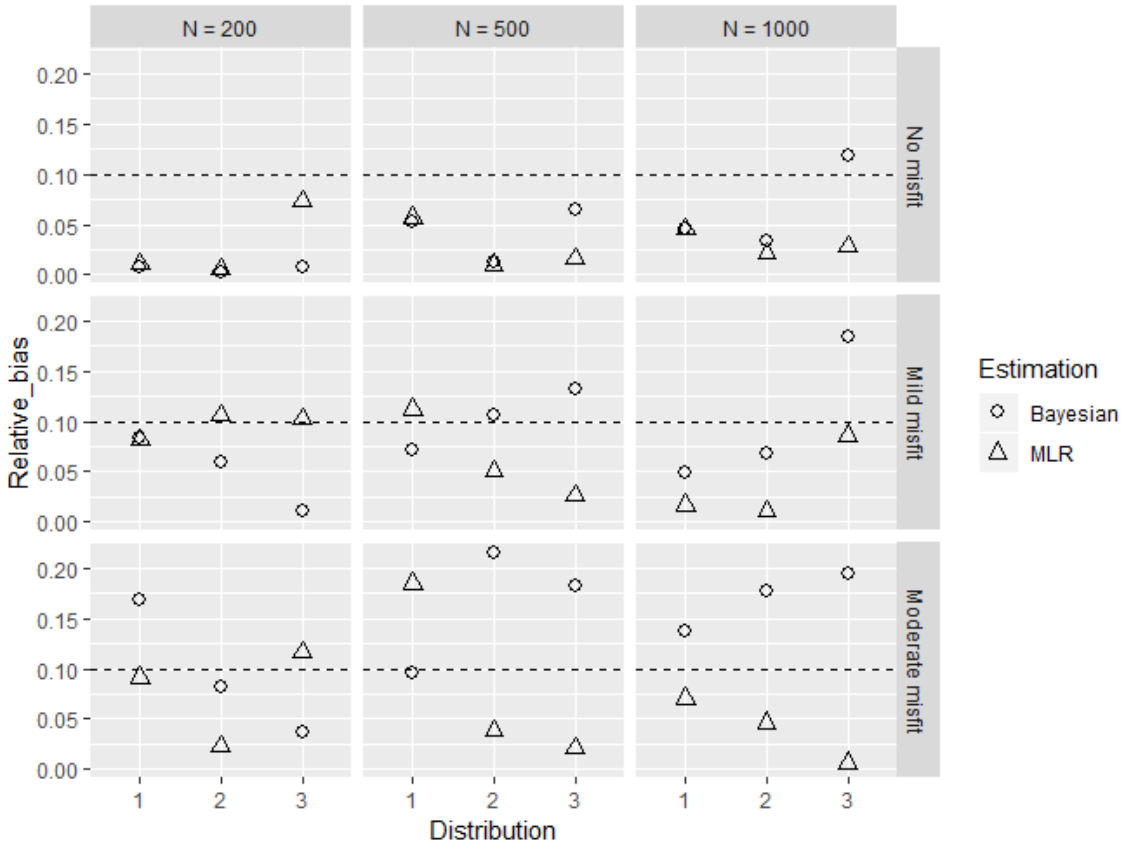


Figure 13. Relative bias of standard errors of inter-factor correlations associated with MLR and Bayesian across Conditions. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Table 43 presents the ANOVA results for relative bias of standard errors of inter-factor correlations estimated using MLR method. It can be seen that the main effects and interaction effects exerted little contribution, with the largest η^2 being .021. The results estimated using Bayesian method are shown in Table 44. Model misfit served as the most noteworthy predictor, accounting for 30.6 % of the variation in the outcome (η^2 = .306). Tukey's HSD multiple comparison results (see Table 45) reveal that relative bias of standard errors increased with degree of model misfit. Moreover, the interaction between

sample size and data nonnormality accounted for 8.5% of the variation in the outcome (η^2 = 0.085). A further inspection of the interaction plot (see Figure 14) reveals that when data were moderately nonnormal, relative bias of standard errors increased with sample size, when data were mildly nonnormal, relative bias increased with sample size then decreased, and when data were normally distributed, relative bias decreased with sample size then leveled off. Other predictors also contributed small proportion of variation, post-hoc comparisons were not discussed here given the complexity of the model.

Comparing MLR and Bayesian estimation methods, Bayesian method yielded smaller overall variation in relative bias of standard errors (Type III SS = 26.242) than that of MLR method (Type III SS = 413.735). Results associated with Bayesian method were mainly affected by model misfit, together with other main effects and interactions, while results associated with MLR did not appear to be significantly affected by the design factors. The results suggest that relative bias of standard errors of inter-factor correlations estimated by Bayesian method are more sensitive to design factors compared to those estimated by MLR method.

Table 43 *ANOVA Results for Relative Bias of Standard Errors of Inter-Factor Correlations with MLR*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	8.791	2	4.395	59.448	<.001	0.021
N	5.439	2	2.719	36.780	<.001	0.013
dist	1.989	2	0.995	13.453	<.001	0.005
cov * N	0.416	4	0.104	1.407	0.229	0.001
cov * dist	0.282	4	0.070	0.953	0.432	0.001
N * dist	1.797	4	0.449	6.075	<.001	0.004
cov * N * dist	0.651	8	0.081	1.100	0.360	0.002
Total	413.735					

Table 44 *ANOVA Results for Relative Bias of Standard Errors of Inter-Factor Correlations with Bayesian Estimator*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
cov	8.037	2	4.019	1992.281	<.001	0.306
N	0.801	2	0.401	198.620	<.001	0.031
dist	1.623	2	0.811	402.265	<.001	0.062
cov * N	0.418	4	0.104	51.753	<.001	0.016
cov * dist	0.714	4	0.178	88.466	<.001	0.027
N * dist	2.228	4	0.557	276.103	<.001	0.085
cov * N * dist	1.663	8	0.208	103.042	<.001	0.063
Total	26.242					

Table 45 *Tukey HSD Multiple Comparisons for the Effect of Model Misfit on Relative Bias of Standard Errors of Inter-Factor Correlations with Bayesian Estimator*

(I) cov	(J) cov	I-J	*SE*	*p*	95% Confidence Interval	
					Lower	Upper
a	b	-0.035	0.002	<.001	-0.038	-0.031
	c	-0.095	0.002	<.001	-0.098	-0.091
b	c	-0.060	0.002	<.001	-0.063	-0.056

Note. cov: model misfit, a = no misfit, b = mild misfit, c = moderate misfit.

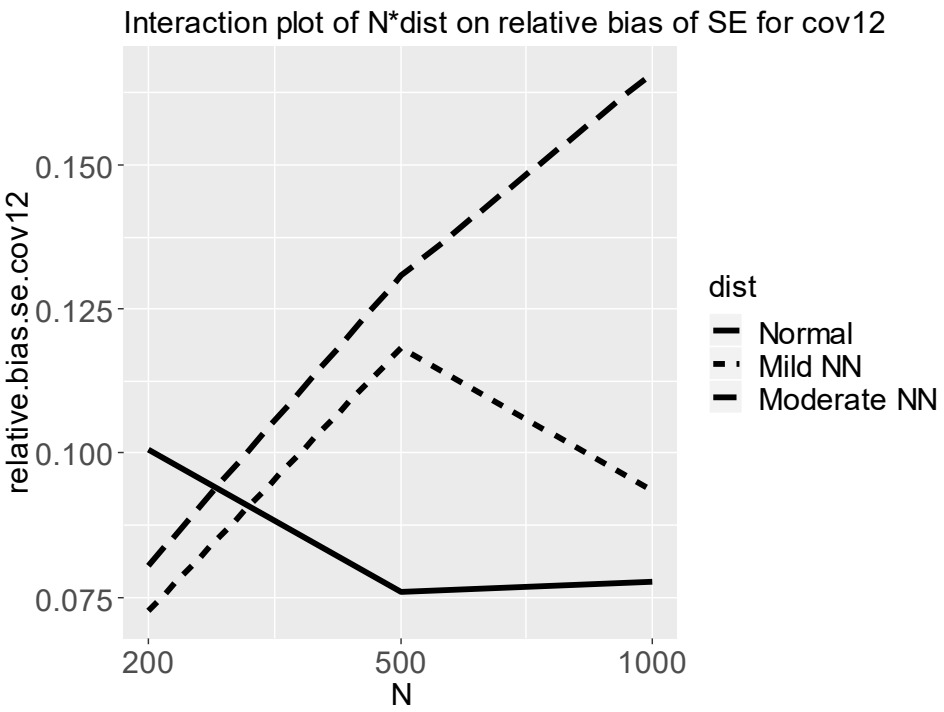


Figure 14. Interaction plot of sample size * distribution for relative bias of standard errors of inter-factor correlation using Bayesian. *Note*. Distribution: 1 = normal, 2 = mild nonnormal, 3 = moderate nonnormal.

Mixed-design ANOVA Results

In order to statistically test the differences between MLR and Bayesian methods, a series of mixed-design ANOVA models were performed for the relative biases of point estimates and standard errors separately. In each ANOVA model, the design factors of sample size, model misfit, data distribution, and their interactions served as between-subjects variables, estimation method and its interactions with design factors served as within-subjects variables. The practical significance was indicated by η^2. Above and beyond the main effect of estimation method, the effects with η^2 larger than or close to .10 were considered as noteworthy.

Table 46 shows the mixed-design ANOVA results for relative bias of loadings. The effects with noticeable contribution were highlighted in grey. For within-subjects effects, estimation method contributed a small proportion of variation in the outcome (η^2 = .068), indicating that there exist differences between MLR and Bayesian estimation methods, through of small magnitude. As can be told from Figure 8, on conditions where method differences were shown, Bayesian method yielded slightly smaller standard errors compared to MLR method. More importantly, the interaction between estimation method and data distribution served as a noteworthy predictor, accounting for 12.4% of the variation. A further inspection of Figure 8 reveals that the differences between the two estimation methods were more salient when the degree of nonnormality was moderate. The remaining within-subjects variables only exerted a small contribution. For between-subjects effects, data distribution accounted for a notable proportion of variation (η^2 = .485). The other between-subjects variables only had small amount of contribution. The between-subjects effects results were consistent with those in the aforementioned separate factorial ANOVA models, and thus post-hoc multiple comparisons were not conducted here. The results suggest that parameter accuracy of loadings associated with Bayesian method are slightly better than those estimated with MLR method, especially when data distribution is moderately nonnormal and sample size is small (based on evidence from Table 46 and Figure 8).

Table 46 *Mixed-design ANOVA Results for Relative Bias of Loadings*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	0.104	1	0.104	495.624	<.001	0.068
estimator * cov	0.001	2	0.000	2.199	0.111	0.001
estimator * N	0.058	2	0.029	137.141	<.001	0.037
estimator * dist	0.191	2	0.095	454.697	<.001	0.124
estimator * cov * N	0.014	4	0.004	16.877	<.001	0.009
estimator * cov * dist	0.001	4	0.000	0.832	0.504	0.000
estimator * N * dist	0.050	4	0.012	59.218	<.001	0.032
estimator * cov * N * dist	0.001	8	0.000	0.331	0.955	0.000
Tests of Between-Subjects Effects						
cov	0.395	2	0.197	17.403	<.001	0.003
N	1.823	2	0.912	80.341	<.001	0.015
dist	59.849	2	29.924	2637.336	<.001	0.485
cov * N	0.052	4	0.013	1.143	0.334	0.000
cov * dist	0.028	4	0.007	0.613	0.654	0.000
N * dist	0.739	4	0.185	16.282	<.001	0.006
cov * N * dist	0.028	8	0.004	0.309	0.963	0.000

Table 47 shows the mixed-design ANOVA results for relative bias of inter-factor correlations. The effects with noticeable contribution are highlighted in grey. For within-subjects effects, all the variables contributed small amount of variation in the outcome (e.g., the greatest contribution was only 2.0%). Regarding the effect of estimation method which is major interest, it had tiny practical significance (η^2 = .006), indicating that there is merely a difference between MLR and Bayesian methods. As can be observed from Figure 9, no salient difference between the two methods was found. The result also echoed the findings in the aforementioned ANOVA models performed separately for the two estimation methods. For between-subjects effects, data distribution was the most

notable predictor, accounting for 26.3% of the variation in the outcome (η^2 = .263). Other between-subjects variables only had small amount of contribution. The between-subjects effects results were also consistent with those in the separate ANOVA models, and thus post-hoc multiple comparisons were not conducted here. The results suggest that relative bias of inter-factor correlations associated with Bayesian method are comparable to those estimated with MLR method.

Table 47 *Mixed-design ANOVA Results for Relative Bias of Inter-Factor Correlations*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	0.036	1	0.036	36.010	<.001	0.006
estimator * cov	0.104	2	0.052	52.238	<.001	0.018
estimator * N	0.115	2	0.057	57.752	<.001	0.020
estimator * dist	0.047	2	0.023	23.418	<.001	0.008
estimator * cov * N	0.060	4	0.015	15.191	<.001	0.011
estimator * cov * dist	0.026	4	0.006	6.522	<.001	0.005
estimator * N * dist	0.002	4	0.001	0.514	0.726	0.000
estimator * cov * N * dist	0.005	8	0.001	0.612	0.769	0.001
Tests of Between-Subjects Effects						
cov	1.579	2	0.790	10.319	<.001	0.003
N	32.524	2	16.262	212.484	<.001	0.054
dist	159.314	2	79.657	1040.823	<.001	0.263
cov * N	0.828	4	0.207	2.706	0.029	0.001
cov * dist	1.315	4	0.329	4.297	0.002	0.002
N * dist	1.529	4	0.382	4.996	0.001	0.003
cov * N * dist	0.729	8	0.091	1.190	0.300	0.001

Table 48 shows the mixed-design ANOVA results for standard errors of loadings. The effects with noticeable contribution were highlighted in grey. For within-subjects

effects, estimation method contributed the largest proportion of variation in the outcome (η^2 = .046), though not of a considerable magnitude, indicating that there exists significant differences between MLR and Bayesian estimation methods. As can be told from Figure 10, Bayesian method yielded smaller standard errors compared to MLR method. The remaining within-subjects variables only exerted small contribution. For between-subjects effects, sample size accounted for the majority of variation (η^2 = .300), followed by data distribution which accounted for 18.6% of the variation (η^2 = .186). The between-subjects effects results were highly consistent with those in the separate ANOVA models performed for MLR and Bayesian methods, and thus post-hoc multiple comparisons were not conducted here. The results suggest that standard errors of loadings associated with Bayesian method are smaller than those estimated with MLR method, though with a small magnitude of effect size.

Table 48 *Mixed-design ANOVA Results for Standard Errors of Loadings*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	0.278	1	0.278	273.706	<.001	0.046
estimator * cov	0.096	2	0.048	47.185	<.001	0.016
estimator * N	0.048	2	0.024	23.471	<.001	0.008
estimator * dist	0.134	2	0.067	66.069	<.001	0.022
estimator * cov * N	0.025	4	0.006	6.193	<.001	0.004
estimator * cov * dist	0.000	4	0.000	0.101	0.982	0.000
estimator * N * dist	0.044	4	0.011	10.926	<.001	0.007
estimator * cov * N * dist	0.011	8	0.001	1.395	0.193	0.002
Tests of Between-Subjects Effects						
cov	0.260	2	0.130	98.161	<.001	0.017
N	4.585	2	2.293	1731.811	<.001	0.300
dist	2.838	2	1.419	1071.869	<.001	0.186
cov * N	0.052	4	0.013	9.872	<.001	0.003
cov * dist	0.004	4	0.001	0.680	0.605	0.000
N * dist	0.482	4	0.120	90.971	<.001	0.031
cov * N * dist	0.013	8	0.002	1.253	0.263	0.001

Note. The highlighted cells represent noteworthy effects.

Table 49 shows the mixed-design ANOVA results for standard errors of inter-factor correlations. The effects with noticeable contribution are highlighted in grey. For within-subjects effects, estimation method contributed the largest proportion of variation in the outcome (η^2 = .174), indicating that there exists significant differences between MLR and Bayesian estimation methods. As can be told from Figure 11, Bayesian method yielded smaller standard errors compared to MLR method. The remaining within-subjects variables only exerted small contribution. For between-subjects effects, sample size accounted for the majority of variation (η^2 = .598), followed by data distribution which accounted for 15.8% of the variation (η^2 = .158). The between-subjects effects results

were highly consistent with those in the ANOVA models (e.g., Table 35, Table 36) built separately for MLR and Bayesian methods, and thus post-hoc multiple comparisons were not conducted for the current model. The results confirm that standard errors of inter-factor correlations associated with Bayesian method are significantly smaller than those estimated with MLR method.

Table 49 *Mixed-design ANOVA Results for Standard Errors of Inter-Factor Correlations*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	0.407	1	0.407	1385.773	<.001	0.174
estimator * cov	0.120	2	0.060	205.141	<.001	0.052
estimator * N	0.069	2	0.034	116.966	<.001	0.029
estimator * dist	0.119	2	0.060	203.019	<.001	0.051
estimator * cov * N	0.006	4	0.002	5.187	<.001	0.003
estimator * cov * dist	0.001	4	0.000	0.542	0.705	0.000
estimator * N * dist	0.049	4	0.012	41.508	<.001	0.021
estimator * cov * N * dist	0.002	8	0.000	0.694	0.697	0.001
Tests of Between-Subjects Effects						
cov	0.245	2	0.123	285.965	<.001	0.022
N	6.791	2	3.395	7918.953	<.001	0.598
dist	1.794	2	0.897	2091.574	<.001	0.158
cov * N	0.010	4	0.002	5.792	<.001	0.001
cov * dist	0.002	4	0.001	1.207	0.305	0.000
N * dist	0.233	4	0.058	135.968	<.001	0.021
cov * N * dist	0.002	8	0.000	0.627	0.756	0.000

Note. The highlighted cells represent noteworthy effects.

Table 50 shows the mixed-design ANOVA results for relative bias in standard errors of loadings. For within-subjects effects, none of the main effects or interaction terms served as a noticeable predictor, with the largest η^2 being .010. Estimation method only contributed a small proportion of variation in the outcome (η^2 = .002), indicating that there exists no salient differences between MLR and Bayesian estimation methods. For between-subjects effects, model misfit was the most noteworthy predictor, accounting for 6.4% of the variation (η^2 = .064). The between-subjects effects results overlapped with those in the ANOVA models built separately for MLR and Bayesian methods, and thus post-hoc multiple comparisons were not conducted for the current

model. The results suggest that relative bias in standard errors of loadings associated with Bayesian method do not differ significantly from those estimated with MLR method.

Table 50 *Mixed-design ANOVA Results for Relative Bias of Standard Errors of Loadings*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	1.597	1	1.597	11.621	0.001	0.002
estimator * cov	7.234	2	3.617	26.327	<.001	0.009
estimator * N	7.564	2	3.782	27.526	<.001	0.010
estimator * dist	4.530	2	2.265	16.486	<.001	0.006
estimator * cov * N	2.342	4	0.586	4.262	0.002	0.003
estimator * cov * dist	1.296	4	0.324	2.358	0.051	0.002
estimator * N * dist	3.623	4	0.906	6.592	<.001	0.005
estimator * cov * N * dist	1.634	8	0.204	1.487	0.156	0.002
Tests of Between-Subjects Effects						
cov	53.465	2	26.733	187.789	<.001	0.064
N	6.648	2	3.324	23.349	<.001	0.008
dist	8.989	2	4.495	31.573	<.001	0.011
cov * N	0.902	4	0.226	1.584	0.175	0.001
cov * dist	5.101	4	1.275	8.958	<.001	0.006
N * dist	1.134	4	0.283	1.991	0.093	0.001
cov * N * dist	2.956	8	0.369	2.596	0.008	0.004

Note. The highlighted cell represents relatively noteworthy effects.

Table 51 shows the mixed-design ANOVA results for relative bias in standard errors of inter-factor correlations. For within-subjects effects, none of the main effects or interaction terms exerted noticeable contribution, with the largest η^2 being .024. Estimation method only contributed a small proportion of variation in the outcome (η^2 = .006), indicating that there exist no remarkable differences between MLR and Bayesian

estimation methods. For between-subjects effects, model misfit served as the most noteworthy predictor, accounting for 7.4% of the variation ($\eta^2 = .074$). The between-subjects effects results were consistent with those in the ANOVA models built separately for MLR and Bayesian methods, and thus post-hoc multiple comparisons were not conducted for the current model. The results suggest that relative bias in standard errors of inter-factor correlations associated with Bayesian method do not differ significantly from those estimated with MLR method.

Table 51 *Mixed-design ANOVA Results for Relative Bias of Standard Errors of Inter-Factor Correlations*

Source	*SS*	*df*	*MS*	*F*	*p*	η^2
Tests of Within-Subjects Effects						
estimator	1.320	1	1.320	34.945	<.001	0.006
estimator * cov	0.079	2	0.039	1.041	0.353	0.000
estimator * N	5.024	2	2.512	66.518	<.001	0.024
estimator * dist	0.075	2	0.037	0.989	0.372	0.000
estimator * cov * N	0.329	4	0.082	2.177	0.069	0.002
estimator * cov * dist	0.359	4	0.090	2.376	0.050	0.002
estimator * N * dist	3.520	4	0.880	23.303	<.001	0.016
estimator * cov * N * dist	1.339	8	0.167	4.433	<.001	0.006
Tests of Between-Subjects Effects						
cov	16.749	2	8.375	219.284	<.001	0.074
N	1.216	2	0.608	15.923	<.001	0.005
dist	3.537	2	1.769	46.313	<.001	0.016
cov * N	0.505	4	0.126	3.305	0.010	0.002
cov * dist	0.637	4	0.159	4.168	0.002	0.003
N * dist	0.505	4	0.126	3.303	0.010	0.002
cov * N * dist	0.974	8	0.122	3.188	0.001	0.004

Note. The highlighted cell represents relatively noteworthy effects

Chapter 5. Discussion

This chapter is a summary of the results, which are discussed and compared to the findings from previous research, suggestions for future directions, and recommendations for applied users.

Summary of the results

Sensitivity of Model Fit Indices

The sensitivity of model fit indices was examined through a series of ANOVA models to inspect the effects of sample size, model misfit, data distribution, and their interactions on each fit index. The ANOVA model for scaled χ^2 shows that the index increases with sample size and model misfit, and decreases with data nonnormality. Moreover, the effects of sample size and nonnormality are more salient under moderate misfit. The results suggest that under misspecified models, the power of scaled χ^2 can be inflated by large sample sizes, and be deflated by nonnormal data distribution. The ANOVA models for scaled CFI and scaled TLI show that these two indices are sensitive to model misfit only, suggesting decent performance. The result for scaled RMSEA shows that it increases with model misfit, decreases with nonnormality, and the effect of data distribution is more salient under misspecified models. A similar pattern is found with SRMR, which is also sensitive to sample sizes. In summary, scaled CFI and scaled

TLI display adequate performance in sensitivity, scaled RMSEA has impaired statistical power under misspecified models due to data distribution, and scaled χ^2 and SRMR are sensitive to sample sizes and data distribution in addition to model misfit.

The inflation of scaled χ^2 towards larger sample size is within expection given that sample size is an embedded parameter in the formula giving rise to χ^2. A possible reason for scaled χ^2 and scaled RMSEA to have impaired statistical power in the presense of data nonnormality is that the adjustment for nonnormality designed in the robust version of χ^2 leads to overcorrection, so that higher level of nonnormality is accompanied by greater adjustment. Scaled CFI and scaled TLI, although as a function of χ^2, are not as sensitive to nonnormality, possibly because their formulas involve a ratio structure of χ^2 for the target model divided by χ^2 for the null model, partly canceling out the weakness of χ^2.

Differences between MLR and Bayesian Methods

In order to contrast MLR and Bayesian estimation methods regarding model fit, rejection rates ($\alpha = .05$) of *p* values associated with the two methods are investigated across the design factors. Under correctly specified models, PP *p* values associated with Bayesian estimation constantly yield rejection rates below .05, and are stable across sample sizes and data distributions. In contrast, *p* values associated with MLR yield rejection rates constantly larger than those of the Bayesian method. Especially, when $N = 200$, MLR over-rejects correct models as the degree of nonnormality increases. With larger sample sizes, MLR produces rejection rates higher than or close to the alpha level of .05. Under mild misfit, both methods start to produce rejection rates larger than or close to .80 when $N = 500, 1000$, and both have impaired statistical power under

moderate nonnormality. Still, the Bayesian method has lower rejection rates than MLR. Under moderate misfit, both methods produce statistical power above .80 regardless of sample sizes. When $N = 200$ and 500, statistical powers are impaired by nonnormality. Altogether, when a model is correctly specified, the Bayesian method has better type I error control under small sample sizes. When there is mild misfit, the Bayesian method is less sensitive to the acceptable degree of misfit than MLR. When there is moderate misfit, Bayesian and MLR have comparable performance when the data are not considerably deviating from normality.

In order to contrast MLR and Bayesian estimation methods regarding parameter estimation, biases of point estimates are investigated first. The inspection of descriptive plots and ANOVA model results reveal that the two estimation methods produce comparable parameter accuracy for both loading and factor correlation estimates. Data nonnormality noticeably increases biases across conditions, while other design factors do not have considerable effects on biases. For loading estimates, relative biases are well controlled below or close to 10% when the data is not deviating from normality much. For factor correlation estimates, a reasonable amount of bias is reached only when $N = 1000$. In summary, when the data is close to a normal distribution, both estimation methods exhibit adequate parameter accuracy across sample sizes for loading estimates, but only under large sample size ($N = 1000$) for factor correlation estimates.

Standard errors for the point estimates are also investigated. For both estimation methods, standard errors decrease with sample sizes, which is expected. Additionally, standard errors increase when the data are moderately nonnormal. For both loading and

factor correlation estimates, Bayesian and MLR methods produce similar results with the exception that under the combined conditions of small sample size and moderate model misfit, the Bayesian method yields lower standard errors than MLR, and is slightly less sensitive to data nonnormality.

The bias analysis results of standard errors show that MLR is robust to sample size, nonnormality, and model misfit, and yields relative bias below or close to 10% in almost all the combined conditions. In contrast, the Bayesian method is sensitive to model misfit and data nonnormality, and produces relative biases above 10% in many conditions especially when the model is moderately misspecified and data is considerably deviating from normality. The results suggest that while Bayesian and MLR produce comparable point estimates, MLR is more robust to all the design factors in terms of bias of standard errors, possibly because the design of robust version of MLE is intended to adjust standard errors for nonnormality.

Comparisons with previous findings

Sensitivity of Model Fit Indices

The current findings regarding scaled χ^2 are consistent with previous research. For example, Jackson (2007) reported that model misfit, and the interaction between misfit and sample size affects χ^2 with noticeable practical significance. Curran, West and Finch (1996) found that the statistical power of scaled χ^2 to reject misspecified models decreased as data nonnormality increased, with a possible explanation that the χ^2 statistics was over-adjusted to accommodate nonnormality. The current study also found that

scaled RMSEA decreases with nonnormality, which is consistent with Nevitt and Hancock's (2000), and Yu's (2002) findings. For the results of SRMR, Yu (2002) had similar finding in the sense that SRMR is deteriorating with sample size, which is supported in the current study. They further reported that SRMR increases with nonnormality, which is not observed in this study. Beauducel and Wittmann (2005) also reported that SRMR is affected by sample size. Moreover, in a principal component analysis performed for multiple model fit indices, they found that RMSEA, SRMR, and χ^2/df loaded on a same component, suggesting that these three indices share common characteristics. This is consistent with our finding in the sense that scaled χ^2, scaled RMSEA, and SRMR are sensitive to design factors above and beyond model misfit, while scaled CFI and TLI are only sensitive to model misfit.

Differences between MLR and Bayesian Methods

Regarding rejection rates of *p* values associated with MLR, Xia, Yung and Zhang (2016) reported that under correctly specified models, the rates are all above or close to .05 regardless of sample size, indicating inflated χ^2 statistics; under moderate misfit, the rates are all above 80%, indicating an adequate statistical power. Their results are supported by the current study. However, in Xia et al.'s (2016) study, the rejection rates are not affected by data nonnormality, which is different from this study. Concerning the performance of MLR associated χ^2 under small sizes, this study shows an over-rejection pattern when $N = 200$, which is consistent with Hu, Bentler and Kano's (1992) finding. Regarding rejection rates of PP *p* values associated with the Bayesian method, Muthen and Asparouhov (2012) conducted simulation studies to compare Bayesian and MLE.

They reported that given no model misfit, MLE inflated χ^2 while PP *p* values had decent type I error control; under ignorable degree of misfit, PP *p* values were less sensitive to the misspecification; under moderate model misfit, PP *p* values had sufficient statistical power to reject the models. These findings are supported by the current study that PP *p* values generally yield rejection rates lower than those of MLR, and under moderate misfit, both have decent statistical power. This study extends Muthen and Asparouhov's (2012) study by incorporating data nonnormality as one of the design factors. Similarly, Liang and Yang (2016) conducted simulation studies to compare Bayesian and MLR methods, and found that rejection rates of PP *p* values are more conservative than *p* values of MLR, which is also observed in this study. They also reported an increase in rejection rates for both methods with degrees of data nonnormality, which is different from the current study.

In terms of parameter estimation performance associated with MLR, Xia et al. (2016) conducted simulation studies based on a CFA model and reported that standard errors for loading estimates are not affected by degree of model misfit, decrease with sample size, and increase with degree of kurtosis. Our findings are consistent with their research in the sense that standard errors are mainly affected by sample size and data distribution, not saliently by model misfit. Regarding the bias of standard errors, under the sample sizes of 200, 500, and 1000, the estimated and empirical standard errors in Xia et al.'s (2016) are comparable when kurtosis = 7, which is also observed in the current study that MLR associated relative bias of standard errors are robust to nonnormality. In terms of parameter estimation associated with the Bayesian method, Muthen and

Asparouhov (2012) compared point estimates of loadings and factor correlations between Bayesian and MLE, and reported that the two estimation methods exhibited comparable performance and both produced little bias in estimation, and thus no preference need to be made between the estimation methods. Our study is consistent with Muthen and Asparouhov's (2012) in the sense that for the point estimates of loadings across model misfit conditions, both estimation methods had similar performance and produced relative biases below 10% when data are not considerably deviating from normality. Different from their results, the relative biases for factor correlation estimates in this study exceeded 10% under smaller sample sizes ($N < 1000$). Moreover, Liang and Yang (2016) compared parameter estimation performance between Bayesian and MLR through simulation studies varying model misfit and data distribution. In their findings, the two estimation methods produced comparable estimates for both loadings and factor correlations, and data nonnormality slightly impaired parameter accuracy. Additionally, the two methods also yielded similar standard errors, while the Bayesian method was more robust to data nonnormality. The results of this study are consistent with Liang and Yang's (2016) in the sense that Bayesian and MLR have comparable point estimation performance, and estimation bias increases under moderate nonnormality. Different from their results which show that point estimates were acceptable across conditions, this study shows that point estimates for factor correlations have considerable bias when sample sizes are smaller than 1000, which is consistent with Lee and Song's (2004) findings that Bayesian method yielded less good performance in inter-factor correlation estimation compared to factor loading estimation.

Suggestions for future directions

Suggestions for future research are made as follows:

First, within the same CFA model, varying magnitudes of model parameters are suggested to be specified to examine the performance of Bayesian and MLR under different loadings and factor correlations. Previous simulation studies (e.g., McNeish, An, & Hancock, 2017) have suggested that holding other conditions constant, increasing degrees of factor loading magnitude (or measurement reliability) is accompanied by worsening RMSEA (higher values of RMSEA) even when the model is properly specified. Moreover, since incremental fit indices (CFI, TLI) indicate the extent to which the target model outperforms the independent model, larger factor loadings may contribute to a better performance of CFI and TLI. In real-world settings, although researchers typically desire for a measurement model with higher reliablity, it is realistic to encounter models with lower magnitude of factor loadings. Therefore, it would be informative to investigate model fit performance and parameter estimation under CFA models with varying degrees of measurement quality (e.g., factor loadings of 0.4 – 0.8).

Second, it is suggested to extend this study to CFA models with greater model complexity. For example, Kenny and McCoach's (2003) simulation study showed that increasing number of indicators in a measurement model is associated with better performance of RMSEA and worsening performance of CFI and TLI. Hence, it would facilitate understandings about model performance by inspecting CFA models with varying number of indicators (or degree of freedom).

Third, in the current study where a noninformative normal prior distribution is specified for Bayesian estimation, future research is encouraged to try different priors such as the *t* distribution to better match the nonnormal data distribution.

Fourth, since this study shows that biases of parameter estimation increase noticeably under moderate data nonnormality, more degrees of nonnormality need to be examined in between the mild and moderate levels, and different combinations of skewness and kurtosis need to be investigated to render understandings of the separate effects from skewness and kurtosis.

Recommendations for applied users

In terms of MLR associated model fit indices, given that scaled χ^2, scaled RMSEA, and SRMR are sensitive to design factors above and beyond model misfit, while scaled CFI and scaled TLI are only sensitive to misfit which is desirable, it is recommended that users rely on multiple fit indices to evaluate performance of a model. More importantly, instead of making decisions simply based on cut-off points of the fit indices, applied users are encouraged to develop a better understanding and interpretation of multiple fit indices combined with a careful inspection of data characteristics such as sample size and nonnormality.

In respect of *p* values associated with Bayesian and MLR methods, given that PP *p* values are robust to small sample size and data nonnormality under correctly specified models, less sensitive to models with ignorable degree of misfit, and have sufficient

statistical power to reject moderately misspecified models, PP *p* values are recommended under small sample sizes.

In terms of parameter estimation associated with the two methods, given that when the data distribution is close to normal, both methods produce acceptable point estimates for loadings regardless of sample size and model misfit, no preference is rendered.

In terms of standard errors, given that the relative biases associated with MLR are robust to data nonnormality, sample size, and model misfit, while Bayesian method is sensitive to both model misfit and data distribution, MLR estimation is suggested for an adequate estimation of standard errors.

References

Asparouhov, T. & Muthen, B. (2010). Bayesian analysis of latent variable models using Mplus. Technical report. Los Angeles: Muthen & Muthen.www.statmodel.com

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin,107,* 238-246.

Bentler, P.M. (1995). *EQS structural equations program manual.* Encino, CA: Multivariate Software.

Beauducel, A. & Wittmann, W. W. (2005). Simulation Study on Fit Indexes in CFA Based on Data with Slightly Distorted Simple Structure. *Structural Equation Modeling, 12*(1), 41-75.

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal* of *Mathematical and Statistical Psychology, 37,* 62-83

Browne, M. W,, & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

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

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29.

Cudeck, R., & Browne, M. W. (1992). Constructing a covariance matrix that yields a specified minimizer and a specified minimum discrepancy function value. *Psychometrika, 57*, 357–369.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.) Hillsdale, NJ: Erlbaum.

Chou, C. P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. in R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Chou, C. P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: A Monte Carlo study. *British Journal of Mathematical and Statistical Psychology, 44*(2), 347-357.

Fan, X. T., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(3), 343-367.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*(1), 56–83.

Fan, X., Felsovalyi, A., Sivo, S. A., & Keenan, S. (2002). *SAS for Monte Carlo studies: A guide for quantitative researchers.* Cary, NC: SAS Institute.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521–532.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Gerbing, D., W., & Anderson, J. C. (2016). Monte Carlo evaluations of goodness of fit indices for structural equation models. *Sociological Methods & Research, 21*(2), 132-160.

Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112*(2), 351-362.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*(4), 424.

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

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

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 650–673). New York, NY: Guilford Press.

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural equation modelling, 10*(3), 333-351.

Lai, K. K. (2018). In Kelley, K., The MBESS R Package. Retrieved from URL http://nd.edu/~kkelley/site/MBESS.html

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39*(4), 653-686.

Lei, M., & Lomax, R. G. (2005). The effect of varying degrees of nonnormality in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 12*(1), 1-27.

Liang, X., & Yang, Y. (2016). Confirmatory factor analysis under violations of structural and distributional assumptions: A comparison of robust Maximum likelihood and Bayesian estimation methods. *Journal of Psychological Science (Chinese), 39*(5), 1256-1267.

Levy, R. (2016). Advances in Bayesian modeling in educational research. *Educational Psychologist, 51*(3), 368-380.

Marsh, H.W., Hau, K-T., & Wen, Z. (2004). In search of golden rules: comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers

in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*(3), 320–341.

McNeish, D., An, J. & Hancock, G. R. (2017). The thorny relation between measurement quality and fit index cutoffs in latent variable models. *Journal of Personality Assessment*, 1–10. https://doi.org/10.1080/00223891.2017.1281286

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105* (1), 156–166.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139.

Merkle, E. C., & Rosseel, Y. (2015). blavaan: Bayesian structural equation models via parameter expansion. Retrieved from http://arxiv.org/ abs/1511.05604

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

Nevitt, J., & Hancock, G R. (2000). Improving the root mean square error of approximation for nonnormal conditions in structural equation modeling. *Journal of Experimental Education, 68*, 251-268.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.

Schumacker, R. E., & Lomax, R. G. (2016). *A beginner's guide to structural equation modeling* (4th ed.). New York, NY: Routledge.

Satorra, A. (1990). Robustness issues in structural equation modeling: a review of recent developments. *Quality & Quantity, 24*, 367-386.

Satorra, A., & Bentler, P. M. (1994). Corrections for test statistics and standard errors in covariance structure analysis. In A. Von Eye, & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research (pp. 399–419)*. Thousand oaks, CA: Sage.

Steiger, H. H., & Lind, J. M. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

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

Van, S. 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), 217-239.

West, S.G., Finch, J.F. and Curran, P.J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In: Hoyle, R.H., Ed., *Structural Equation Modeling: Concepts, Issues, and Applications*, Sage, Thousand Oaks, 56-75.

Xia, Y., Yung, Y. F., & Zhang, W. (2016). Evaluating the Selection of Normal-Theory Weight Matrices in the Satorra–Bentler Correction of Chi-Square and Standard Errors. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(4), 585-594.

Yu, C.Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. (Unpublished doctoral dissertation). University of California, Los Angeles.

Yang, Y., & Liang, X. (2013). Confirmatory factor analysis under violations of distributional and structural assumptions. *International Journal of Quantitative Research in Education,1*(1), 61-84.