Posterior Predictive Model Checking of Local Misfit for Bayesian Confirmatory Factor Analysis

Chi Hang Au

Follow this and additional works at: https://commons.lib.jmu.edu/master201019

Part of the Educational Assessment, Evaluation, and Research Commons, and the Quantitative Psychology Commons

Recommended Citation

Au, Chi Hang, "Posterior Predictive Model Checking of Local Misfit for Bayesian Confirmatory Factor Analysis" (2018). Masters Theses. 543.
https://commons.libjmu.edu/master201019/543
Posterior Predictive Model Checking of Local Misfit for Bayesian Confirmatory Factor Analysis

Chi Hang Au

A thesis submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

For the degree of

Masters of Arts

Department of Graduate Psychology

August 2018

FACULTY COMMITTEE:

Committee Chair: Allison J. Ames, Ph. D.

Committee Members/Readers:

Sara J. Finney, Ph. D.

Deborah L. Bandalos, Ph. D.
Acknowledgement

First, I would like to thank my advisor, Dr. Allison Ames, for her wonderful support, patience, and expertise on my thesis. This project was a huge undertaking. Without her, I would not be able to finish this project. Also, the feedback from my thesis committee was crucial in making this thesis as rigorous and presentable as possible. Therefore, thank you to Sara and Debbi!

Second, I would also thank my fantastically brilliant CARS friends Nick, Derek, and Aaron. Talking with you all made me a better researcher and provided the much-needed encouragement and push for me to succeed. I always learn something new whenever I am hanging out with you all whether it be validity arguments, SAS tricks, or just random, but fun, world/life advices.

Lastly, I would also like to thank everyone that had helped me along the way but that I couldn’t think of when I am writing this acknowledgement. I am incredibly grateful for everything.
# Table of Contents

Acknowledgements .......................................................................................................................... ii  
Table of Contents .............................................................................................................................. iii  
List of Tables ........................................................................................................................................ vi  
List of Figures ........................................................................................................................................ vii  
Abstract ............................................................................................................................................... ix  
Chapter I: A Bayesian Cenrury ................................................................................................................. 1  
  Technical Reasons ................................................................................................................................. 3  
  Philosophical Reasons ......................................................................................................................... 4  
  Key Concepts in Bayesian Statistical Modeling ............................................................................... 5  
  Model-Data Fit .................................................................................................................................... 7  
  Model-Data Fit Reporting ............................................................................................................... 8  
  Modeling for Psychoeducational Instrument ............................................................................ 12  
  Research Questions ......................................................................................................................... 15  
  Organization of Study ....................................................................................................................... 15  
Chapter II: Review of Literature ............................................................................................................. 17  
  Analytic Framework ............................................................................................................................ 18  
  Estimation Approaches: Overview ............................................................................................... 19  
  Inference and Estimation ................................................................................................................... 20  
  Current Issue in Model Evaluation ............................................................................................... 20  
  Overview of Bayesian Inference .................................................................................................... 21  
  Prior Information on Parameters of Interest ................................................................................ 22  
    Types of Prior Specifications ........................................................................................................ 22  
    Examples ........................................................................................................................................ 25  
  Data Likelihood ............................................................................................................................... 26  
  Posterior Distribution ....................................................................................................................... 28  
  Bayesian Structural Equation Modeling ..................................................................................... 30  
  Estimation ......................................................................................................................................... 31  
    Considerations of Estimation Methods ..................................................................................... 32  
  Benefits to Bayesian Estimation to Modeling ............................................................................ 34  
  Statistical Advantages ..................................................................................................................... 34
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior Parameter Estimate Comparisons</td>
<td>94</td>
</tr>
<tr>
<td>Data Validation</td>
<td>96</td>
</tr>
<tr>
<td>Visual Comparison of Heatmaps</td>
<td>98</td>
</tr>
<tr>
<td>Congruence</td>
<td>100</td>
</tr>
<tr>
<td>Summary</td>
<td>105</td>
</tr>
<tr>
<td>Chapter V: Discussion</td>
<td>106</td>
</tr>
<tr>
<td>Summary of Results</td>
<td>106</td>
</tr>
<tr>
<td>Global Fit</td>
<td>106</td>
</tr>
<tr>
<td>Research Question 1</td>
<td>107</td>
</tr>
<tr>
<td>Research Question 2</td>
<td>107</td>
</tr>
<tr>
<td>Discussion of Findings</td>
<td>108</td>
</tr>
<tr>
<td>Significance of Findings</td>
<td>112</td>
</tr>
<tr>
<td>Limitations</td>
<td>113</td>
</tr>
<tr>
<td>Directions for Future Research</td>
<td>114</td>
</tr>
<tr>
<td>Conclusion</td>
<td>115</td>
</tr>
<tr>
<td>Appendix A</td>
<td>116</td>
</tr>
<tr>
<td>Appendix B</td>
<td>117</td>
</tr>
<tr>
<td>Appendix C</td>
<td>118</td>
</tr>
<tr>
<td>Appendix D</td>
<td>119</td>
</tr>
<tr>
<td>Appendix E</td>
<td>120</td>
</tr>
<tr>
<td>Appendix F</td>
<td>125</td>
</tr>
<tr>
<td>References</td>
<td>130</td>
</tr>
</tbody>
</table>
List of Tables

Table 1. Comparisons between Frequentist and Bayesian Paradigm. ...............................5
Table 2. Key Terms and Definitions.........................................................................................6
Table 3. Fit Evaluation Procedures for Confirmatory Factor Analysis Used in Literature. 8
Table 4. Reporting Guidelines Differences...............................................................................11
Table 5. Previous BCFA Applications.......................................................................................44
Table 6. Commonly Reported Fit Indices..................................................................................50
Table 7. Studies on Behaviors and Use of Fit Indices...............................................................52
Table 8. Descriptive Statistics for FA16 SOS...........................................................................65
Table 9. One-Factor Model Unstandardized and Standardized Coefficients and Standard
        Errors.................................................................................................................................71
Table 10. Two-Factor Model Unstandardized and Standardized Coefficients and Standard
        Errors....................................................................................................................................72
Table 11. One-Factor Bayesian $ppp$ Fit Conclusion Matrix (.025-.975) ......................101
Table 12. One-Factor Bayesian $ppp$ Fit Conclusion Matrix (.05-.95) .........................101
Table 13. One-Factor Bayesian $ppp$ Fit Conclusion Matrix (.10-.90) .........................102
Table 14. Two-Factor Bayesian $ppp$ Fit Conclusion Matrix (.025-.975) ......................102
Table 15. Two-Factor Bayesian $ppp$ Fit Conclusion Matrix (.05-.95) .........................102
Table 16. Two-Factor Bayesian $ppp$ Fit Conclusion Matrix (.10-.90) .........................103
Table 17. Number of Misfitting Item-Pairs Flagged.............................................................104
Table 18. Disagreement of Fit Conclusion Between Methods..............................................104
Table 19. Skewness, Kurtosis, and Mardia's Normalized Multivariate Kurtosis of
        Observed Data and Posterior Predictive Datasets..........................................................110
List of Figures

Figure 1. Path diagram example.................................................................12

Figure 2. Latent factor relation for the SOS..................................................28

Figure 3. Posterior predictive model checking procedure for local fit evaluation.......67

Figure 4. One-factor structure for the SOS....................................................68

Figure 5. Correlated Two-factor structure for the SOS....................................69

Figure 6. Example of a traceplot that suggests convergence has achieved...........78

Figure 7. Example of a traceplot that suggests convergence has not achieved.........79

Figure 8. Traceplots for the one-factor model.............................................89

Figure 9. Traceplots for the two-factor model.............................................90

Figure 10. Posterior density plots for the one-factor model.............................91

Figure 11. Posterior density plots for the two-factor model............................92

Figure 12. Autocorrelation plots for the one-factor model...............................93

Figure 13. Autocorrelation plots for the two-factor model...............................94

Figure 14. One-factor model estimates comparison of the two estimation frameworks.95

Figure 15. Two-factor model estimates comparison of the two estimation frameworks.96

Figure 16. Scatterplot of “true” and recovered parameter values for the one-factor model.
....................................................................................................................97

Figure 17. Scatterplot of “true” and recovered parameter values for the two-factor model.
....................................................................................................................98

Figure 18. Heatmaps containing ppp and Frequentist correlation residuals for the one-factor model
....................................................................................................................99
Figure 19. Heatmaps containing \textit{ppp} and Frequentist correlation residuals for the two-factor model.
Abstract

Posterior predictive model checks (PPMC) are one Bayesian model-data fit approach. Thus far, PPMC for Confirmatory Factor Analytic applications focused primarily on global fit evaluation, ignoring the nuanced information in local misfit diagnostics. This study developed a PPMC approach for local misfit and applied it to a test-taking motivation scale. If the PPMC approach is effective, fit conclusions derived from the PPMC approach should be congruent with the fit conclusions derived from the Frequentist approach. Number of item-pairs flagged as misfitting and number of disagreements were computed to evaluate congruence. Congruence is achieved if the number of item-pairs flagged as misfitting is equivalent under the Frequentist and the Bayesian approach and the number of disagreements is zero. Although congruence was not achieved, the present research sets up a foundation for future research in local fit evaluation using PPMC.
CHAPTER I
A BAYESIAN CENTURY

A recent systematic review of Bayesian articles in psychology (van de Schoot et al., 2017) investigated Krushke’s 2011 claim that “the 21st century is becoming Bayesian.” The use of Bayesian methods in applied psychological work has steadily increased since the 1990s and is growing rapidly in the early 2000s (van de Schoot et al., 2017). Bayesian methods are an unconventional estimation and inferential framework for statistical analyses. This method incorporates prior information to inform estimation and emphasizes the plausibility of parameter estimates. The Bayesian approach is not limited to one subfield or discipline of psychology, but is used in a wide variety of contexts. The reasons for choosing a Bayesian approach are equally as diverse, with oft-given reasons being: “the use of priors, estimating otherwise intractable models, circumventing small sample size issue, and so forth” (van de Schoot et al., p. 223). As such, the application of Bayesian methods and inferences became a topic worthy of investigation.

Part of the reason for the growth in Bayesian methods can be found in Wainer (2010), who said that Bayesian methods are a suite of tools researchers must have in order to successfully tackle research problems looming in the future. The Bayesian approach allows researchers to use previous knowledge to inform the plausibility of relations between variables. He says, “Bayesian methods allow us to do easily what would be hard otherwise,” continuing, “[and] facility with them is a must for anyone who intends to make contributions to measurement in the future” (p. 7). Models that involve many variables and parameters are difficult to analyze and interpret under the conventional Maximum Likelihood approach, which Bayesian methods can help overcome. Further, it has been suggested that researchers should become familiar with
not just the terms and broad concepts of Bayesian methods, but that topics related to prior distributions [the distribution of plausible values] and the Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm should become second nature (Wainer, 2010).

Although there are many advantages to the Bayesian approach, gaining mastery in this framework is not easily realized in practice. Van de Schoot and coauthors (2017) outlines several reasons. First, traditional statistical training (i.e., the Frequentist paradigm) teaches psychologists in terms of null hypothesis significance testing. Van de Schoot and coauthors lament that “Psychologists have largely been married to their null hypotheses, and they often design studies in terms of a point-null hypothesis.” They also say that the use of Bayesian methods for the uninitiated researcher is not user-friendly, and that accessible software is not available to psychologists. Thus, education and exposure are the main impediments related to the use of Bayesian statistics in psychology.

In their systematic review, Rens van de Schoot and coauthors (2017) found that the most common Bayesian publication topic used regression methods (including Item Response Theory models and Structural Equation Models). More specifically, structural equation modeling (SEM) was the second-most common type of Bayesian statistical model used in psychology. SEM was a close second to the use of Bayesian IRT models. Moreover, there has been a steady increase in Bayesian applications over time, with a dramatic spike in 2012 (van de Schoot et al., 2017). One reason may be that, to assist practitioners who are interested in using Bayesian methods, statistical packages is becoming increasingly easy to use (Levy, 2009). The authors of the review predict more and more Bayesian structural equation modeling (BSEM) in the coming years as
statistical packages for Bayesian estimation become more user-friendly and accessible (van de Schoot et al., 2017).

There are two main categories for the appeal of Bayesian methods: technical and philosophical. Each has contributed to the growth of Bayesian methods described in van de Schoot, et al. (2017) and Levy (2009). A description of each is described below.

**Technical Reasons**

Part of the reason for Bayesian structural equation modeling (BSEM) growth has been the implementation of the Bayesian estimator in the software Mplus (Muthén & Muthén, 1998-2012), initially described in Muthén (2010). As reviewed by van de Schoot and coauthors (2017), software for Bayesian estimation has become more user-friendly and accessible for practitioners in psychology. Researchers justified the use of Bayesian methods to resolve issues regarding features of data such as assumption violations, intractable models, and model identifications (van de Schoot, 2017), that may pose problems in conventional software estimation packages if conventional estimation were employed. Particularly, Mplus has been a popular statistical package, accounting for 20% of Bayesian publication in the reviewed journals in 2015. The benefit of Mplus is its ease of Bayesian estimation capability. Unlike other statistical packages, syntax for Bayesian analysis merely adds on a few extra lines of codes that specify the choice of estimator, types of priors, and number of iterations for basic modeling needs. The downside to this simplicity is its lack of flexible options.

Other programs such as the SAS MCMC procedure and Stan (Stan Development Team, 2017) provide more options and user control for Bayesian estimation. This increased flexibility requires the user to specify more detailed model statements and
manage multiple script files. However, R packages such as blavaan (Merkle & Rosseel, 2015) have shown to be promising for future applications due to its accessibility (it is a free statistical package capable of implementing various procedures using simple syntax) and similarity to other popular R packages.

**Philosophical Reasons**

Another reason for the growing popularity is the inference that Bayesian studies can offer. Because parameters (the quantifications of relations between test items and latent constructs) are conceptualized as random variables in Bayesian framework, their values can take on a range of numeric values. As such, by incorporating prior information gained outside of the study when evaluating observations, inferences can be made about the parameters directly using posterior distributions (a product of prior information and the observed data) without referencing the commonly used $p$-values for statistical significance and unobserved null distributions. Further, the heart of Bayesian inference is updating beliefs about parameters in light of evidence. Under this framework, posterior distributions from one study can be incorporated into future studies via prior specification to form a continuous chain of updating knowledge, allowing results from previous study to inform parameter estimation. This function of Bayesian analyses could also mitigate the reproducibility crisis in psychology (Etz & Vandekerckhove, 2016). These advantages will be better elaborated in the literature review. Therefore, both the advent of accessible Bayesian estimation software and philosophical attractiveness of Bayesian inferences led to the steady rise of Bayesian publications.

To contrast the difference between Frequentist paradigm and Bayesian paradigm, Table 1 provides a brief comparison of key features between the two.
Table 1
Comparisons between Frequentist and Bayesian Paradigm

<table>
<thead>
<tr>
<th></th>
<th>Frequentist</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical Theory</td>
<td>Relies on large-sample statistical theories, theoretical distributions.</td>
<td>Does not rely on statistical theory.</td>
</tr>
<tr>
<td>p value</td>
<td>Probability of obtaining the observed statistics or more extreme assuming the null hypothesis is true.</td>
<td>Probability of a hypothesis being true.</td>
</tr>
<tr>
<td>Parameter Estimate</td>
<td>Returns a set of parameter estimates that maximizes the likelihood of the observed data. Assumes parameter values are constants fixed in the population.</td>
<td>Returns a distribution of plausible parameter estimates. Assumes parameter values are random variables that can take on a range of values in the population.</td>
</tr>
<tr>
<td>Model Fit Conclusion</td>
<td>A model is either plausible or not plausible in producing the observed data</td>
<td>Models are viewed in terms of plausibility. How well the data supports the model?</td>
</tr>
</tbody>
</table>

Key Concepts in Bayesian Statistical Modeling

Before introducing the purpose of the present research, it is necessary to clarify the terminology that will be used throughout the manuscript. Bayesian statistical modeling refers to the approach of approximating data generating mechanisms (that cannot be directly observed) using prior information, the data likelihood, and the estimated posterior distribution(s). This approximation of data generating mechanism should be consistent with the observed data and should be able to explain the relations among observed variables. In order to describe and understand the data generating mechanism, prior information (information derived from theory before observing the data
that quantifies uncertainty or certainty about parameter values) is used in conjunction with observed data to make inferences about parameters that are the cogs in the data generating mechanism. The results of Bayesian analyses are posterior distributions of plausible parameter values, which are obtainable by combining prior information and observed data likelihood through the use of Bayes’ theorem. To summarize relevant key terms, refer to Table 2.

Table 2  
*Key Terms and Definitions*

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistical model</td>
<td>Description of relations between variables such as latent constructs, observed item responses.</td>
</tr>
<tr>
<td>Data generating mechanism</td>
<td>The unobserved statistical model underlying reality that produces the observed data.</td>
</tr>
<tr>
<td>Parameters</td>
<td>Coefficients in the model that describe specific relations or information about variables.</td>
</tr>
<tr>
<td>Estimation</td>
<td>Computation methods used to obtain numerical variables for parameters within a model.</td>
</tr>
<tr>
<td>Prior information</td>
<td>Knowledge about parameters obtained outside the observed data (e.g., expert opinion, substantive theory)</td>
</tr>
<tr>
<td>Data likelihood</td>
<td>Numeric expression of the extent of support a sample of data can provide for a particular set of parameter solution.</td>
</tr>
<tr>
<td>Posterior distribution</td>
<td>Range of plausible parameter values.</td>
</tr>
<tr>
<td>Model-Data fit</td>
<td>Procedure used to evaluate whether the estimated model adequately reproduced the observed relations</td>
</tr>
<tr>
<td>Global fit</td>
<td>Type of model-data fit information that summarizes reproduced relations into single value.</td>
</tr>
<tr>
<td>Local fit</td>
<td>Type of model-data fit information that provides specific reproduced relations between variables.</td>
</tr>
</tbody>
</table>
Table 2
*Key Terms and Definitions – Continued*

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posterior Predictive Model Checking</td>
<td>One approach in Bayesian method to evaluate model-data fit. Uses posterior predictive data to gauge how well the model reproduced features from the observed data to plausible, future data.</td>
</tr>
<tr>
<td>Posterior Predictive Data</td>
<td>Simulated data from the posterior distributions of parameters assuming the estimated model is indeed the true data generating mechanism.</td>
</tr>
<tr>
<td>Discrepancy Measure</td>
<td>Feature of data or test statistics to evaluate within the observed data and the posterior predictive data.</td>
</tr>
<tr>
<td>Posterior Predictive P value</td>
<td>Proportion of posterior predictive data that generated discrepancy measure value greater than or equal to the observed data value. Used in posterior predictive model checking.</td>
</tr>
<tr>
<td>Heatmap</td>
<td>Data visualization method to visually communicate direction and magnitude of a set of numeric values.</td>
</tr>
</tbody>
</table>

**Model-Data Fit**

Because Bayesian methods emphasize the plausibility or uncertainty of parameter values, this method requires a different strategy for model-data fit from Frequentist fit indices. Model-data fit evaluation indicates how well the theoretical model represents the observed data. In order to calculate model-data fit, Gelman and coauthors (2004) recommend posterior predictive checking (PPMC), which uses test statistics (e.g., a chi-square value) to compute posterior predictive values (ppp values). That is, if the model is the data generating mechanism for the observed data, simulated data (posterior predictive data) from the model should share similar features with the observed data (e.g., the chi-square value should be similar). The discrepancies between the observed data, specified model, and posterior predictive data are summarized via ppp values. The ppp value is defined as the proportion of the posterior predictive data that has the test statistics
exceeding or equal to the test statistics value in the observed dataset. This *ppp* is then compared to a cutoff (conventionally .05-.95) to make decisions about model-data fit. Commonly, *ppp* value of .5 is considered good fit and values outside .05-.95 are considered misfit. For example, the observed discrepancy between a model and the observed data is described by a $\chi^2$ value of 200. Using the same model, posterior predictive data are simulated by randomly sampling from the posterior distributions of parameters and randomly proposing factor scores. Then, the same model is fit to each of the posterior predictive dataset and their $\chi^2$ values computed. If only 3% of these $\chi^2$ values are larger than 200 (the observed value), the *ppp* value is .03. This would be interpreted as global model misfit, meaning the model did not adequately reproduced observed relations in the data. A brief listing of fit evaluation procedures used under Frequentist and Bayesian paradigm are listed in Table 3.

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Fit Evaluation Procedures for Confirmatory Factor Analysis Used in Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Global Fit</td>
</tr>
<tr>
<td>Frequentist Paradigm</td>
<td>$\chi^2$ (Classical Likelihood Ratio)</td>
</tr>
<tr>
<td></td>
<td>RMSEA</td>
</tr>
<tr>
<td></td>
<td>SRMR</td>
</tr>
<tr>
<td></td>
<td>CFI</td>
</tr>
<tr>
<td></td>
<td>Various Fit Indices</td>
</tr>
<tr>
<td>Bayesian Paradigm</td>
<td>$ppp_{\chi^2}$</td>
</tr>
<tr>
<td></td>
<td>$ppp_{SRMR}$</td>
</tr>
</tbody>
</table>

**Model-Data Fit Reporting**

Although model-data fit provides crucial information to consider before interpreting results, van de Schoot and coauthors (2017) highlighted a lack of
comprehensive reporting of model-data fit for Bayesian models. *P* values were reported in only 32 articles (19.2%; including SEM, IRT, ANOVA models) reviewed by van de Schoot and coauthors (2017). *P* value reporting is depended on whether the statistical packaged used provides this information by default (Mplus). For example, 40.9% of the reviewed BSEM articles reported *pp* values, whereas 96% of the articles using analysis of variance did not report the *pp* value, which indicates the lack of consistent model-data fit evaluation procedures across statistical models. As stated by van de Schoot and coauthors, “[o]verall, the posterior predictive *p* value is not used as a standard tool to evaluate the fit of a model.” Further, of the seven BSEM applications presented on the Mplus website, none investigated local misfit using *pp* values for BSEM models. That is, none of the seven BSEM applications examined fit at the item level under PPMC framework. Only global fit is investigated for these studies. As such, flexible options for *pp* value computations are lacking for fit evaluation on procedures under the Bayesian framework, which serves to motivate the present study. As there are various fit evaluation procedures available for Frequentist SEM studies, so too should there be various options to evaluate model-data fit for Bayesian studies.

Although van de Schoot and coauthors (2017) examined *pp* value reporting in their review of Bayesian publications, there was no mention of local fit evaluation. That is, information regarding whether the model adequately reproduces observed relations at the item level was not examined. As the goal of model fit evaluation is to obtain evidence to assess whether the specified model adequately represents the data, it is necessary to explore misfit at both the global (overall model-data fit) and local (particular relationships reproduced by model) levels. Under the Frequentist framework, global fit
results can mask local misspecification, making it necessary to evaluate fit using residual correlations. The residual correlation matrix is obtained by computing the difference between the observed correlation matrix and the model-implied correlation matrix. Model misspecification would be manifested as large residuals and/or clustering of sizable residuals. Under the Bayesian framework, global fit is typically evaluated using the $ppp$ value of the $\chi^2$ statistics (classical likelihood ratio). If this $ppp$ value lies outside .05-.95, there is model misfit. Similar to Frequentist framework, this $ppp$ value does not provide information at the local level. However, local fit is seldom discussed, as will be shown in the subsequent literature review. When local fit is examined, researchers typically estimate error covariance matrices in the model and check their posterior distributions to determine if the matrix contains zero. If that parameter contains zero and that the plausibility of this zero value is considered high, it may suggest local misfit. However, contrary to the prevailing use of $ppp$ values for global fit evaluation in Bayesian analyses, only Levy (2011) has used the posterior predictive checking approach for local fit evaluation using inter-item correlations. Therefore, further investigation into local fit using $ppp$ values is necessary to provide practitioners with options to evaluate local fit consistent with the typical Bayesian approach.

Another impetus for expanding current model fit evaluation is the publication and reporting of results. The natural end-point of any scientific endeavor should be dissemination of study findings to accumulate knowledge and to promote replications. Therefore, it is necessary to set guidelines in reporting standards to document crucial and necessary information sufficient for criticism and replication. Recently, the American Psychological Association (APA) published new reporting standards containing a section
devoted to Bayesian analysis (Appelbaum et al., 2018). This is yet another indication that Bayesian studies are becoming common-place in psychology and related fields, which warrants dedicated attention to setting reporting standards.

For Frequentist SEM studies, the APA standards explicitly require a discussion of local fit evaluation involving the residual correlation matrix. In contrast, for Bayesian studies, the reporting standards for model fit are far less detailed. Table 4 illustrates the difference in detailed guidance on model fit evaluation for the two estimation approaches from the APA guidelines (Appelbaum et al., 2018). This comparison further demonstrates the lack of consistent fit evaluation guidelines and the need for accessible procedures for practitioners.

Table 4

<table>
<thead>
<tr>
<th>Reporting Guidelines Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approach</td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td>Frequentist SEM</td>
</tr>
<tr>
<td>Bayesian</td>
</tr>
</tbody>
</table>
Modeling for Psychoeducational Instrument

In the context of psychometric evaluation of psychoeducational instruments, the hypothesized data generating mechanism is tested via a series of steps to obtain support for particular score interpretations (e.g., appropriateness of subscale score interpretation). This general statement relates to factor analytic models in particular. First, a model is specified to describe the theoretical relations among unobservable variables (latent factors) and observed variables (indicators). The model describes the process by which the observed data is generated. This step is often facilitated through the drawing of a path diagram, such as the one found in Figure 1. Each arrow (or path) in the diagram represents one type of parameters to be estimated (e.g., factor pattern coefficient, interfactor covariance, error covariance). These parameters are quantifications of the relations between the observed (\(X_1\) through \(X_{10}\) in Figure 1) and unobserved variables (\(L_1\) and \(L_2\) in Figure 1). Moreover, other types of parameters such as error variances will also be estimated.

![Path diagram example](image)

*Figure 1. Path diagram example*

Under the Bayesian approach, before evaluating the observed data, each parameter specified in the model is assigned a prior distribution. Prior distributions
(priors) can be specified in a number of ways, such as through previous study findings, content experts, and other sources of evidence. For example, $\text{Normal}(0, sd=10)$ represents that a parameter is normally distributed around a mean of zero with a standard deviation of ten (i.e., we expect the parameter to not be significantly different from zero; that is, it is expected that the most plausible value for this parameter is zero but the standard deviation of ten represents uncertainty). The mean of a prior distribution specification indicates the most likely value that a parameter can take on, whereas the standard deviation captures how certain/uncertain the modeler is regarding the center of the prior.

After priors are specified for each parameter, the model is estimated from the observed data to obtain posterior distributions of the plausible values of each parameter by combining priors with the observed data likelihood. The data likelihood is a measure of the extent to which an observed sample of the data provides support for particular values of a parameter in a parametric model. Bayesian estimation of the parameter values is typically carried out using some type of sampling algorithm, such as Markov chain Monte Carlo (MCMC) (more on this in Chapter III). However, before interpreting the posterior distributions of parameters, the model must fit the data to demonstrate the appropriateness of the results.

Ignoring local misfit may pose threats to validity when making inferences. Local misfit may occur due to unrealistic assumptions imposed on the model (e.g., fixing certain parameters to zero when they should be free to vary) or a new factor may need to be included in the model. Therefore, checking for local misfit may provide insight into misfit resulting from negative wording effects and item clustering, and these insights can be used to inform further model specification. Moreover, parameter estimates from a
misfitting model may be biased, which will undermine the validity of inferences. As such, it is always necessary to evaluate model-data fit to gain deeper understanding of the appropriateness of the inferences made from a model.

A model-data fit procedure, therefore, is an exercise in checking whether the model adequately represents the data. Under the Frequentist approach, a model-implied covariance matrix is created from the parameter estimates in factor analytic models. If a model was misspecified, the model’s parameter estimates would not be able to adequately reproduce the relations among observed variables (i.e., the model-implied covariance matrix based on parameter estimates would have different values than the observed covariance matrix). One approach to fit evaluation under the Bayesian framework is through *posterior predictive model checking* (PPMC; Gelman et al., 2004). Under this approach, future, plausible observations/datasets (posterior predictive data) are generated from the posterior. If the model is a good fit, features of posterior predictive data should resemble those same features of the observed data. Although model-fit evaluation is critical to consider before interpreting the parameter estimates, procedures under the Bayesian framework lack specific guidelines in understanding local misfit.

In sum, there is a need to expand options to evaluate local misfit. The purpose of this study is to evaluate the efficacy of PPMC and the related posterior predictive *p* value (*ppp* value) approach for factor analytic models as one way to evaluate local misfit. Specifically, the study will use inter-item correlation residuals. In other words, because the posterior predictive data are simulated from the specified model, the posterior predictive data should fit (align with) the model better than the observed data. Hence the discrepancy measure (correlation residuals) should be larger between the observed data
and the model versus the posterior predictive data and the model. This study will
investigate whether the \textit{ppp} approach detect this misfit. Three different cutoff criteria
were used to flag misfit in this study (.05-.95, .1-.9, and .15-.85) to challenge
conventional cutoff criteria of .05-.95. \textit{Ppp} values that lie outside these intervals will be
considered model-data misfit as the posterior predictive data are considered discrepant
from the observed data. Further, heat maps will be used to visually facilitate
interpretation of this information for instruments with many items. One observed dataset
was used in this study to demonstrate the efficacy of this approach by specifying two
models: one model that is supported by previous literature and one model that is
unsubstantiated (worse-fitting model). The research questions are articulated in the
following.

\textbf{Research Questions}

RQ1. Do posterior predictive \textit{p} values, using inter-item correlation residuals as
the discrepancy statistics, provide congruent local fit conclusions with Frequentist
estimation for Bayesian Confirmatory Factor Analysis model?

RQ2. Which cutoff criteria for posterior predictive \textit{p} values will lead to most
congruent conclusions between the Frequentist and Bayesian approach to local fit
evaluation?

\textbf{Organization of the Study}

To set up a foundational understanding of the current study, an overview of
relevant literature on Bayesian inference, estimation, and fit evaluation will be presented
in Chapter II before a discussion of methodology in Chapter III. Additionally, Chapter II
will provide a comparison of Frequentist and Bayesian estimation methods, an
introduction to Bayesian structural equation modeling, and a discussion of the Bayesian fit evaluation procedure using posterior predictive model checking. Chapter III will outline the methodology used to answer the present research questions using data from a psychoeducational instrument. Chapter IV will provide results pertaining to the research questions. Lastly, Chapter V will present discussions, significance, limitations, and directs for future research.
CHAPTER II
REVIEW OF THE LITERATURE

Psychology as a science employs statistical methods to make inferences from observations and extends them to make general claims in the population (Rosnow & Rosenthal, 1989). Specifically, statistical results are used as evidence to champion competing hypotheses derived from different theories (Dienes, 2011). In applied psychological and educational research, for example, tests and instruments are commonly used to gather information about unobservable, latent constructs. These observations are then used to make inferences or claims regarding levels on the unobservable constructs. As such, it is crucial to accumulate evidence supporting the relation between the item and the latent construct to avoid making erroneous conclusions regarding person’s level on the construct.

In instrument development, competing hypotheses are expressed as different configurations of relations among latent and observed variables. Parameters are quantifications of features of a model regarding relations between test items and a latent construct. One particular type of parameter (i.e., factor pattern coefficients and interfactor covariance) is used to make claims about the internal structure of an instrument. Gathering evidence of internal structure provides insight to inform score interpretation; thus, examining internal structure is a crucial step in instrument development. The following introduction presents an overview of the type of evidence needed to inform the inferences from psychological and educational assessments, with a focus on the methods evaluating internal structure. A test-taking motivation example will be used throughout this section to illustrate the various concepts.
Analytic Framework

Confirmatory Factory Analysis (CFA) is a structural equation modeling (SEM) technique used to examine the fit of a theoretical model to the observed data – the pattern of relations between instrument items and the theoretical construct(s) that represents the data – to substantiate particular inferences (Brown, 2006; Kline, 2011). CFA is used to test the fit of a model to the data, which informs a researcher’s interpretation of scores. Consider the Student Opinion Survey (SOS; Sundre & Moore, 2002; Thelk et al., 2009), designed to measure two aspects of motivation when students complete an assessment: perceived importance and examinee effort. Each of the scales is measured by five items. To test this theoretical two-factor model, CFA is used. Evaluation of model-data fit is a critical component to informing the inferences made from SOS data. For example, if a one-factor (single construct) model fits the data as well as than a two-factor model, one would not be justified to compute two subscale scores since there is only one meaningful construct that influences the responses. In contrast, if the two-factor model appears to adequately fit the data and better than a one-factor model, then reporting and making inferences from the two subscales is appropriate.

To select among competing models, model-data fit information is computed using the estimated parameters. There are two approaches to estimating parameters associated with CFA models: Frequentist (e.g., maximum-likelihood approaches) and Bayesian. What follows is a description of these two estimation approaches, fit evaluation procedure appropriate under each approach, and types of appropriate inferences.
**Estimation Approaches: Overview**

The Frequentist approach is commonly taught and adopted by applied researchers. Practitioners following this approach typically use Maximum Likelihood-type estimation methods that maximize the plausibility of data given auditioned parameter values. Under the Frequentist approach, the focus of inference is made from the auditioned, “true”, parameter to the data. Specifically, the plausibility of the data is the emphasis. That is, under this framework, practitioners assume that one set of true parameter solution exists in the population that generated the observed data, and this set of true parameter solution maximizes the likelihood of observing the obtained data. A typical procedure of this estimation is as followed. A set of parameter values are assumed to be true. Then, the likelihood of the observed data is evaluated under that specific parameter solution.

On the other hand, Bayesian methods do not rely on Maximum Likelihood estimation. As such, the inferences made are no longer from the parameters to the data. Instead, the inference is from observed data to the plausible values of the parameter. That is, prior specifications of plausible parameter distributions are combined with the likelihood of the data to construct a set of plausible parameter values. In other words, practitioner’s previous knowledge of the plausible parameter values are updated using the observed data likelihood. This approach has been less commonly utilized in psychology. However, recent availability of analytic packages that permit Bayesian estimation has permitted an increase in adoption of this approach. Inferences made from the data to the parameter(s) are expressed by the posterior distribution that provides a range of plausible parameter values. As such, the two paradigms do not produce the same types of inferences.
**Inference and Estimation**

Recall that users of psychological and educational assessments seek to make general claims using observed data in describing test-takers’ level on unobservable constructs: the level on the construct based on a sample of a person’s behavior. This type of statement resembles the logical structure of inductive reasoning (Godrey-Smith, 2003). In this sense, the objective is to make an inference(s) about the parameter(s), such as the pattern of relations among test items and latent constructs (what are plausible parameter values give the observed data is obtained). This is in contrast to the Frequentist approach that makes inferences about the observed data (how likely the observed data is obtained assuming a set of parameter values is true).

Estimation methods should be reflective of the type of inference one wishes to make. Inductive logic is best reflected in the Bayesian approach; this is the motivation behind the use of Bayesian inference.

**Current Issue in Model Evaluation**

Despite recent developments in Bayesian techniques for CFA models, current analytic options limit evaluation of key aspects of the model – some types of misfit are neglected as a result. One such type of misfit is local misfit. Local misfit refers to the bivariate relations among instrument items that were not adequately reproduced by the model (Kline, 2011). It may be possible to determine that, overall, the model is a good fit to the data despite the presence of local misfit (see, for example, Johnston & Finney, 2010). Thus, ignoring local misfit may threaten the validity of inferences made from the instrument.
Returning to the SOS example, if a few of the SOS items are negatively worded, and this negative wording influences responses but this mechanism is not modeled, local misfit indices (i.e., correlation residuals) should detect such an issue (i.e., should detect model misspecification). Ignoring local misfit would affect estimated relations between latent and observed variables (i.e., parameter estimates would be biased). The present research seeks to contribute to the discourse of Bayesian CFA (BCFA) fit and create an accessible approach to examine and communicate evaluation of local fit.

The following literature review will detail Bayesian inference, the Bayesian Structural Equation Modeling (BSEM) approach, and model evaluation for Bayesian Confirmatory Factor Analysis (BCFA) models using the posterior predictive model checking (PPMC) approach. The goal is to build the case for local fit examination that is lacking in current BCFA applications. To accomplish this goal, the review of literature begins with an overview of Bayesian inference.

**Overview of Bayesian Inference**

In general, Bayesian estimation methods require specification of a likelihood model for the data, as well as specification of parameter priors (Gelman, Carlin, Stern, & Rubin, 2004). After the data is observed, prior information on the parameters is combined with information from the likelihood to provide a posterior distribution of parameter information. This section details the main Bayesian modeling stages, which will be used to facilitate discussion of the literature regarding BCFA and model-data fit in the Bayesian framework.

Bayesian estimation uses three types of information to construct inferences about parameters: the prior distribution, the likelihood, and the posterior distribution. The
following discussion presents an overview of these three components of Bayesian inferences. More comprehensive discussions regarding Bayesian methods can be found in other texts such as *Bayesian Data Analysis* (Gelman et al., 2013), McElreath’s *Statistical ReThinking* (2015), Lynch’s *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists* (2007), Kaplan’s *Bayesian Statistics for the Social Sciences* (2013), and didactic articles such as van de Schoot et al. (2014) and Gelman and Shalizi (2013).

**Prior Information on Parameters of Interest**

The prior distributions of plausible parameter values quantify knowledge and information about the model parameters before observing the data (Kaplan, 2013). Within the Bayesian modeling context, specification of the prior distribution of the parameter (i.e., assigning a normal distribution with a center and spread) serves as an integration of previous research findings, explicitly to affect estimation of parameter values. These prior distributions of the parameter are often selected for their mathematical properties, as well as their informativeness regarding parameter estimates. For example, early applications of Bayesian estimation in factor analytic model used priors specifications of plausible parameter values to avoid theoretically implausible values, such as negative variances (Martin & McDonald, 1975).

**Types of prior specifications.** There are different ways to encode prior knowledge of plausible parameter values for Bayesian modeling (Carlin & Louis, 2000). Priors that contain possible values of the parameter of interest informed by content experts are termed *elicited priors*. Specifically, these priors can be gathered from previous study results. For example, two datasets were independently obtained from two different time points. The posterior of the first dataset can be used as priors to obtain the
posterior of the second dataset (Carlin & Louis, 2000). Examples of integrating prior research can be found in epidemiology (e.g., Hamra, Richardson, Maclehose, & Wing, 2013), but have yet to be implemented in BSEM.

Prior specification could also come from common distributional families, such as the Normal or Gamma distributions, which are tied to distributional assumptions. Distributional families are also used in specific combinations because of the desirable property on conjugacy, that the posterior distribution would share the same distributional form with the prior specification. For example, for binomially-distributed data, a Beta prior will result in a Beta posterior distribution. Conjugacy allows for simple and straightforward computation of the posterior distribution. These are also considered elicited priors, although less subjective than expert-elicited priors.

Typically, the influence of a prior distribution on the results is controlled by the distribution’s variance, with smaller variances demonstrating more strength and prior precision. For instance, a $\text{Normal}(0, 1)$ prior distribution would be considered more precise and stronger than a $\text{Normal}(0, 100)$ prior distribution. Put another way, a larger variance suggests ignorance, whereas a smaller variance reflects more confidence in the center of the prior. If no existing knowledge is available (e.g., exploring new domain), it is still necessary to quantify this ignorance in Bayesian inference. Prior specifications of this type are referred as noninformative or diffuse priors. Although it is commonly called “noninformative,” some prior information is still expressed, such as through specifying a range of values (Kaplan, 2013). For example, a truncated normal distribution may be used that restricts values between -10 and 10, but is considered noninformative in that range (large variability in plausible parameter values within an upper and lower bound).
A common noninformative prior distribution is the uniform distribution. This distribution implies that the parameter estimates should be within a specific range but that each value within that range has equal probability; that is, no parameter value is more likely than the rest. Many Bayesian statistical packages use the uniform distribution as a default prior specification due to several desirable characteristics. For example, Mplus (Muthén & Muthén, 1998-2012) uses a $\text{Normal}(0, +\infty)$ prior, which is effectively uniform, for factor loadings in CFA models. In general, noninformative priors that use a uniform distribution should have less influence than the data likelihood on the posterior distribution as sample size increases (Bauwens, Lubrano & Richards, 2003). Intuitively, this is logical in that ignorance is updated by accumulation of evidence (from the observed data). In fact, priors are often referred to in terms of their “prior sample size,” which is inversely proportional to the prior variance. That is, a large prior variance implies a small prior sample size, reflecting little to no prior information. However, the range of values must be specified carefully so to avoid an improper prior distribution, which may result in an improper posterior distribution. On the other hand, if existing knowledge is available, informative prior distributions should be used.

Informative priors are the strength and one appeal of Bayesian inference. These priors are driven by substantive knowledge, expert opinions, distributional assumptions, and previous data. The center of the prior probability distribution reflects the most plausible value for the unknown parameter, and the degree of certainty for that value can be expressed by a small standard deviation. The choice of distribution can be motivated by known mathematical properties (Kaplan, 2013). For example, when modeling a Bernoulli random variable, the beta distribution is used as a prior for the probability of
success due to several desirable properties: (1) the probability distribution integrates to 1.0 (i.e., a proper prior); (2) when both outcomes are equal, the distribution is symmetric and peaks at .5; and (3) when little prior information is available, the distribution is flat and resembles a uniform distribution.

**Examples.** To demonstrate how, and why, priors are assigned to differing parameter types, turn to a BCFA example described in Levy and Mislevy (2016). A single-factor model is presented, using an example drawn from Mitchell (2009; as described in Levy & Mislevy, 2016). This instrument consists of thirty-one, five-point Likert-type items intend to measure student collegiate experiences with and perception of peers, faculty, intellectual growth, and academic goals (Pascarella & Terenzini, 1980; as described in Levy & Mislevy, 2016). Five subscales are derived by taking the average of the item scores of the items for that subscale, and the subscale scores are used as indicators of the single latent trait.

Diffuse priors were employed for all parameter values to reflect ignorance in possible parameter values. Factor loadings were assigned a common prior of \( \text{Normal}(1, 10) \) to reflect a wide range of possible values centered around 1. The variance of 10 suggests a wide range of plausible values centered around 1. Intercepts were assigned a prior of \( \text{Normal}(3, 10) \) because of “knowledge of the response scale” (Levy & Mislevy, 2016). More specifically, because the latent factor was given a mean of 0, the intercept represents the expected response for an examinee at the mean of the latent factor. Because the items were integer scores from 1 to 5, Levy and Mislevy (2016) felt a centered in the middle at 3 was appropriate. \( \text{Inverse-Gamma}(5, 10) \) is assigned to error variances as their values are positively skewed and unbounded.
Later, Levy and Mislevy (2016) modeled two latent factors, with three observables loading onto one factor, and two loading onto the other. Because the latent factor variances were estimated, one of the factor pattern coefficients was set to one to resolve factor indeterminacy along with fixing the latent variables means to zero. The other four factor pattern coefficients were still assigned priors \( \text{Normal}(1, 10) \). Because factor covariances are typically distributed in an elliptical distribution, a prior of \( \text{Inverse-Gamma}(5, 10) \) is assigned because these values are positively skewed and unbounded. Such a specification reflects beliefs that the latent variables are likely positively correlated (correlation of approximately 0.3), but that considerable uncertainty is present.

To illustrate further the concept of assigning priors, the logic in Levy and Mislevy (2016) is applied to the student SOS example. The factor pattern coefficients (the relations between the items and the test-taking motivation construct) would be assigned a prior of \( \text{Normal}(0, +\infty) \) and the error variances would be assigned \( \text{Inverse-Gamma}(5, 10) \) prior to quantify the uncertainty of these parameters based on the default settings of Mplus. In another example, Muthén and Asparouhov (2012) assign cross-loadings and residual covariances of \( \text{Normal}(0, 0.01) \). Notice the extremely informative prior (the variance of 0.01 suggests a very narrow prior, represents high certainty) that does not fix the cross-loading at zero, but allows for some flexibility to reflect theory that if the latent factors are correlated, the items likely relate to other factors.

**Data Likelihood**

Another aspect of Bayesian inference is the likelihood of the observed data, which represents the same logical convention as in Frequentist estimation. The likelihood principle states that only the information from the data summarized by the likelihood is
considered support of a parameter, serving as the basis of maximum likelihood estimation. That is, the observed data can be used to support a set of parameter values to be true in the population if that set of parameter values maximizes the likelihood of the observed data (Kaplan, 2014). Therefore, the parameter set that has the highest likelihood of generating the observed data is interpreted as candidate of the true parameter value.

In the Bayesian framework, the data are assumed to affect the posterior inference only through the likelihood (Gelman at al., 2004). The likelihood is the observed evidence from the data in light of parameter population values. The conceptualization is the same as it is under the Frequentist framework. In other words, it reflects the likelihood of observing the sample of data given some population parameters. Mathematically, the likelihood is proportional to the conditional probability of the data given the parameter where the proportionality does not depend on the parameter. In the Bayesian framework, this likelihood is used as evidence to update the prior probability distribution to construct the posterior distribution.

The equations that follow detail the multivariate likelihood functions for these CFA models. Equation (1) represents the multivariate normal likelihood function suitable for models involving many variables. However, for the present study, corrections for non-normality of the data are applied to the $\chi^2$ statistics and standard errors in Frequentist estimation. Equation (2) denotes the conditional distribution of the observed data given the parameters.
The likelihood corresponding to Figure 2 is:

\[ L(X|\mu, \Sigma) = \prod_{i=1}^{N} f(x_i) = (2\pi)^{-\frac{NP}{2}}|\Sigma|^{-\frac{N}{2}} \exp \left[ \sum_{i=1}^{N} -\frac{(x_i^T - \mu)^T \Sigma^{-1} (x_i^T - \mu)}{2} \right] \]  

(1)

and the conditional distribution for Figure 2 is:

\[ p(x_i|\Xi, \kappa, \phi, \Lambda, \Psi) = \prod_{i=1}^{n} \prod_{j=1}^{I} p(x_{ij}|\xi_i, \tau_j, \lambda_j, \psi_{jj}) \]  

(2)

In equation (1), \( X \) refers to the observed responses to an item whereas \( \mu \) and \( \Sigma \) refer to the mean and variance of latent variable parameter values respectively. In equation (2), the bolded terms are matrices: \( \Xi \) is the latent variable scores, \( \kappa \) is the mean of the latent variable, \( \phi \) is latent (co)variances, \( \Lambda \) is factor pattern coefficient, \( \Psi \) is the error (co)variances, \( \xi \) is the latent variable scores for examinees, \( \tau \) is the latent variable intercepts, \( \lambda \) is the factor pattern coefficients, and \( \psi \) is the error (co)variances.

**Posterior Distribution**

The final component in Bayesian inference is the posterior distribution, serving as the basis for Bayesian inference. By leveraging the prior with the likelihood, the posterior distribution serves as a compromise between the two, and its features detail the updated range of plausible parameter values and their associated credibility. Often, estimating the posterior is not directly possible via Bayes’ Theorem because of the analytical complexity involved in integrating complex likelihoods. This is the case with CFA
models. Markov Chain Monte Carlo (MCMC), a general-purpose sampling method that relies on the Monte Carlo principle, is used instead for the task of constructing the posterior when the analytical complexity becomes too great for direct computation.

Inferences can be made once the posterior is constructed. From the posterior distribution, central tendency measures, such as the mean and mode, describe the most plausible parameter value. Posterior standard deviations describe the spread of the posterior probability distributions. Subsequently, the highest density interval, or credibility interval, can be constructed to indicate a range of plausible parameter values at a certain degree of credibility. Thus, compared to ML estimation, which provides a point estimate, inference from the posterior distribution provides a description of a range of plausible values. The extent to which inferences made from parameter posteriors are sensitive to model misspecification is an important consideration in a Bayesian analysis (O’Hagan, 1994). Despite construction of the posterior via MCMC, the likelihood component (i.e., the CFA model of interest) must be checked for model misspecification—that is, their adherence to model assumptions needs to be deemed adequate before inferences can be validly drawn.

Put in a larger context, Bayesian inference uses the posterior distribution to reflect the uncertainty of the parameter as opposed to frequentist methods that yield a single point estimate and standard error (to reflect sampling uncertainty) for each parameter. Moreover, this framework is consistent with the scientific goal of accumulation of evidence; the posterior distribution can be used as a prior distribution for future studies.

In sum, Bayesian inference begins with understanding substantive theory to specify appropriate priors to quantify available knowledge or lack thereof. Then, the
likelihood of the data is used to weigh the prior probability distribution to update credibility of the parameter. After the posterior distribution is obtained, plausible values are interpreted. Subsequently, in future studies, features from the previous posterior distribution are used to inform the next prior specification to continue the process. This process of directly incorporating learned knowledge is well suited for the scientific endeavor in accumulating evidence.

At the inference level, Bayesian inference is compatible with making prescriptive statements in applied research (the values of certain quantities given the observed data). In order to make prescriptive statements, factual, descriptive premises are coupled with at least one evaluative premise (Searle, 1964). The purpose of statistics, in science, is to provide evidence to evaluate claims and hypotheses (Sun & Pan, 2011). In this light, statistical decisions made under Frequentist methods are typically dichotomous, such as reject or fail to reject the null hypothesis based on information from a single study. In contrast, Bayesian inference is better at accumulating evidence to update credibility of parameters. Statistical decisions are made in degrees of certainty and that plausibility of models can be examined. As such, the application of Bayesian modeling warrants exploration and is discussed next.

**Bayesian Structural Equation Modeling**

Bayesian Structural Equation Modeling is a family of models that uses Bayesian inference and estimation for structural equations models. All three components from the previous overview of Bayesian inference are included in BSEM. The present discussion will focus on Bayesian Confirmatory Factor Analytic models, which seek to estimate the relation between latent and manifest variables using Bayesian methods.
Early applications of Bayesian inference in factor analytic models include Martin and McDonald (1975), Lee (1981), and Scheines, Hoijtink, and Boomsma (1999). Bayesian inference was first applied to exploratory factor analysis models to alleviate the issue of Heywood cases (i.e., extreme, theoretically implausible values such as negative variances; Martin & McDonald, 1975). Specifically, incorporation of Bayesian prior information was proven possible in Martin and McDonald (1975) and Lee (1981) later generalized the Bayesian approach to confirmatory models.

The benefits that Bayesian statistical inference provides over conventional Maximum Likelihood (ML) estimation were discussed in Scheines, Hoijtink, and Boomsma (1999). However, the benefits of BSEM were not actualized in applied research until the expansion of Markov Chain Monte Carlo (MCMC) estimation methods using popular statistical software such as R (e.g., the package Blavaan, Stan) and Mplus (Muthén & Muthén, 1998-2012). The Bayesian framework offers a different perspective and approach to making inferences from parameter estimates. The following discussion will provide a general overview of Bayesian estimation and the statistical and substantive advantages of this paradigm. Also included is a discussion about the drawbacks of BCFA models.

**Estimation**

Estimation methods for BCFA models are distinct from Frequentist estimation. Frequentist estimation for confirmatory factor analysis, due to its reliance on long-run frequencies, usually rely on Maximum Likelihood procedures. The ML estimation paradigm seeks to maximize the log-likelihood of the data given auditioned parameter values. In contrast, Bayesian estimation uses Markov Chain Monte Carlo (MCMC)
procedures such as the Metropolis-Hasting algorithm or Gibbs sampler to describe and summarize parameter posterior distributions (Levy, 2016).

**Considerations of Estimation Methods.** Maximum Likelihood and Bayesian methods differ in their treatment of prior information and underlying statistical theory. Bayesian estimation methods explicitly quantify prior information into the estimation procedure whereas Frequentist methods treat each study as independent, without input from previous findings. Under the Frequentist approach, parameters are viewed as fixed constants and plausible estimates are obtained by auditioning possible true values. On the other hand, the Bayesian approach treats parameters as random variables that can take on a range of plausible values. The following is a general comparison of the two estimation methods in these domains.

Perhaps the biggest difference between the two estimation approaches is the use of prior information in MCMC procedures, which does not typically occur in ML methods. The priors define a probabilistic model for the model parameters. This is a key distinction between Bayesian and Frequentist paradigms: the Bayesian paradigm is founded on the notion that model parameters can be described by a distribution representing the probability that the parameter equals each possible value.

Prior information can come from many sources. These were briefly discussed in the previous Bayesian overview section. Leamer (1983) classified sources of prior information by degree of confidence. Truths such as axioms merit the highest confidence, followed by data-derived facts, expert opinions, and convention at the lowest level. There is a clear subjective element to using expert-elicited information in the formation of priors. Inherently, prior specification lacks objectivity in the ontological sense in that
previous findings can directly influence inferences drawn in current and future studies. However, this is one of the strengths of Bayesian inference. The prior probability changes as new data are observed, and the credible values of parameter inference are updated accordingly. This is typically referred to as Bayesian updating and is consistent with the goal of accumulation of evidence in science (Jackman, 2009).

The effect of prior information on inference is particularly apparent with small samples. In fact, this is one of the reasons many early adopters turned to Bayesian inference – results can be trusted as they perform well with sparse data and small sample sizes when asymptotic theory is unlikely to hold and Frequentist approaches are limited (Fox, 2010). Another use of priors is that they can help avoid extreme or theoretically impossible values such as negative variances. Recall that the posterior distribution is proportional to the product of the prior probability and the likelihood of the data. The prior probability distribution has more influence on the posterior when there are fewer observations. Conversely, the more observations that are available (i.e., more evidence), the less impact prior beliefs have on posterior inferences. Priors that are informative (i.e., those with smaller variances) affect the posterior distribution more heavily than uninformative priors (i.e., those with larger variances).

Another difference in estimation is that MCMC samplers do not invoke large sample asymptotic theory (Muthén & Asparouhov, 2012). Rather than obtaining parameter estimates using estimators with desirable asymptotic properties that are appropriate to the data type and sampling distribution, MCMC samplers can accommodate a wide range of data types with appropriate prior specification. In sum,
Bayesian estimation methods make less restrictive assumptions than Frequentist estimation methods and incorporate prior information directly into the modeling process.

**Benefits to Bayesian Estimation to Modeling**

There are many decided benefits of the BCFA and BSEM approaches. Although many of these have been alluded to already, this section provides more detail on the benefits of the BCFA. Many benefits relate to relaxing the restrictive assumptions imposed by Frequentist large-sample theory – both statistical and substantive.

**Statistical Advantages**

One of the main advantages of BCFA lies in its use in small sample scenarios. Statistically, Bayesian estimation of measurement models do not use the restrictive assumptions imposed by Frequentist estimators (e.g., forcing secondary factor pattern coefficients as zero). The reliance on large sample asymptotic theory is requisite for popular Maximal Likelihood (ML)-like estimators. When sample size is small, the sampling distribution is unknown, or poorly defined, for parameter estimates. Because fit indices are computed from parameter estimates assumed to be true, the sampling distribution of the parameter estimates and their associated tests will also be unknown. The effects of this interaction of sampling distribution and sample size is problematic when evaluating fit of the models. The result is a potential reduction of the precision of the estimators due to different forms of non-normality that impact standard error, which threatens the validity of inferences (Kline, 2011). It follows that if the standard error is inflated or deflated, significance tests cannot be trusted. For more and more complex models, larger and larger sample sizes are necessary. Thus, the validity of inferences made from the model can be artificially threatened due to sample size. Therefore, the
statistical and practical significance of parameter estimates are questionable in small sample scenarios.

The issue of sample size can be mitigated by the incorporation of priors in the Bayesian framework; information from small-sample data is leveraged against expert knowledge or previous research findings to produce the parameter posterior distributions from which inferences are made. Extreme or theoretically implausible values that arise from likelihood alone are effectively avoided (Martin & McDonald, 1975). When interpreting the credibility of parameter estimates, validity of Bayesian inference is grounded on the characteristics from the posterior distribution rather than the asymptotic behavior of estimators. Because sampling distributions are not involved, plausible parameter values are obtained and applied to prior theory directly. In other words, the observed data are used to update a prior belief; when sample size is small, the prior belief has more influence. The resulting posterior distribution can still provide interpretable parameter estimates.

**Substantive Advantages**

Substantively, liberation from Frequentist estimators permits testing for a greater variety of models that more realistically represent underlying theory. When employing Frequentist estimators, unique solutions cannot be obtained for under-identified models; so to achieve over- and just-identified models, cross-loadings and correlated residuals are usually constrained to zero. At the theoretical level, relationships between variables that are fixed at zero are considered non-existent. These models often fail to account for the complexity of reality and theory. Such models, as Cole, Ciesla, and Steiger (2007) illustrated, ignore known measurement properties for the sake of model identification.
This can bias parameter estimates and possibly negate or discredit the inferences to the theory drawn from the erroneous models. When the estimates are biased, inferences related to the magnitude of relations between variables will suffer.

In contrast, Bayesian methods for measurement models that were under- or just-identified in the Frequentist framework can be compared on model-data fit. These under-identified models estimate cross-loadings and correlated residuals and are often better reflections of reality. Muthén and Asparouhov (2012) formalized a set of procedures for BSEM models that freely estimate cross-loadings and correlated residuals. By specifying near-zero priors on cross-loading and correlated residual paths, Muthén and Asparouhov (2012) demonstrated greater fit for theoretically plausible models than conventionally constrained models while accounting for known measurement properties. Since more information are incorporated in the forms of prior specifications, under- and just-identified models are not serious estimation issues as long as the prior specifications are appropriate.

**Drawbacks of Bayesian Inference in Modeling**

A common criticism for Bayesian inference is the specification of priors. Particularly, the subjective nature of incorporating prior probability appears to be counter-productive to the ideal of scientific objectivity (Berger, 2006). This concern has been repeatedly raised, partly due to the early champions of Bayes methods. These pioneers argued, rather forcefully, that all statistical calculations should be done after one’s prior beliefs on the subject had been carefully evaluated and quantified (Carlin & Louis, 1996). These arguments resulted in apprehension that results could easily be
manipulated by the statistician, research funding body, or bureaucratic entities, leading to conclusions and policies that were not objectively valid.

Recall that Frequentist estimation relies on long-run frequencies, assuming an observed sample is one of infinite samples from the same population that may or may not contain the true, fixed parameter. In this sense, probabilities are tied to the assumed large sample asymptotic behavior of the estimator and not to the parameter of interest. Further, it follows that each resampling for future studies is independent from information learned from previous studies; the incorporation of previous findings is not made explicit. Therefore, probabilities of evidence used to make decisions regarding parameter estimates are unaffected by previous or future research. However, in applications, previous research findings are used to guide future studies and should affect decisions drawn from future observations. Moreover, accumulation of evidence (pooling results from various studies) is necessary for science to grow. Thus, the Frequentist $p$ value is not sufficient for this endeavor. Bayesian inference, on the other hand, subscribes to a different view of probability.

As neatly illustrated by de Finetti (1974), probabilities quantify uncertainty in subjective experience and do not exist independently in reality. Bayesian inference adheres to this definition. In application, conclusion statements for hypothesis testing are inherently probabilistic. As such, hypotheses regarding parameter values should be treated with degrees of credibility, not dichotomous (reject vs. fail to reject) decisions. This treatment is more in line with the desirable endeavor of accumulation of evidence in that previous findings can be incorporated into future studies by quantifying their credibility in the form of prior probabilities. In this framework, credibility of parameter
estimates is updated by future observations. Thus, the subjectivity that is inherent with prior specification can be a boon.

There are also operational disadvantages associated with Bayesian estimation including longer runtime and accessibility of software packages. Similar to Frequentist estimation, MCMC algorithms use an iterative process. Recall that a major difference is that Frequentist estimation yields a single point estimate per parameter whereas Bayesian estimation yields an entire probability distribution of estimate. Complex models require more iterations to converge and with the increase in model-complexity, storage of parameter posteriors faces memory limits on computational platforms. For example, Ames (2015) found run-times of longer than 1 hour for dichotomous 3-PL IRT models using SAS PROC MCMC. No such information has been provided for BCFA models.

Furthermore, convergence, the conclusion of the iterative process, differs between the two approaches. ML-like estimators converge to a solution when the change in likelihood is minimal between iterations. In MCMC, convergence is determined both graphically, and supported statistically using different indices. As such, MCMC has a subjective element to determining whether convergence has occurred.

Another operational drawback is the lack of user-friendly and accessible software packages. Recent software programs capable of handling BCFA include Mplus, SAS, JAGS, and R (i.e., the package Stan, Blavaan). However, Mplus and SAS can be expensive. Open source software such as R can be either restrictive in data type (e.g., the package blavaan cannot handle categorical data) or require more advanced technical knowledge (e.g., Stan requires programming ability in unfamiliar languages). Lack of user-friendly and accessible software may deter applied researchers from exploring
Bayesian modeling. However, these disadvantages may well be temporary with advancements in analytic options.

This section merely scratched the surface of the debate on Bayesian and Frequentist approaches in modeling. A more detailed contrast between Bayesian and Frequentist inference can be found in Wagenmakers et al. (2008), as a start. Nevertheless, despite some drawbacks, the benefits of BCFA and the application of Bayesian methods should not be discounted. One notable drawback of BCFA models, which this study was designed to alleviate, is the lack of diversity in rigorous model fit evaluation approaches. Although models are useful tools to test and understand the internal structure of instruments, estimates are untrustworthy if the model is not a good representation of the data. Therefore, Bayesian approaches to model fit evaluation are addressed next.

**Bayesian Confirmatory Factor Analysis: Methods of Model Fit Evaluation**

No discourse on model estimation is complete without discussion of model fit. For factor analytic models, model-data fit refers to how well the CFA model can reproduce the observed relations in the data. This evaluation can be summarized either as an overall model-fit index, or as a local evaluation which focuses on pairs of variables at a time.

Regardless of choice of estimation method, fit evaluation is necessary in theory-testing and model selection. After all, validity of the inferences using parameter estimates from ill-fitting models are questionable, at best. Compared to the rich body of literature in Frequentist estimation methods regarding fit, this area is still growing in the Bayesian SEM literature. Under the Bayesian modeling approach, Gelman et al. (2004) proposed three approaches to evaluate model-data fit – one of which, Posterior Predictive Model
Checking (PPMC), is the focus of this study. This section first provides a description of PPMC. Then, two types of fit (i.e., global and local) will be discussed in the context to the present research scenario. As such, the present research seeks to contribute to the BSEM literature by expanding analytic options for BCFA-PPMC.

**Posterior predictive Model Checking**

Posterior Predictive Model Checking (PPMC) seeks to compare features of simulated datasets to the observed dataset by taking random draws from the posterior distribution (Levy & Mislevy, 2016). Recall that the posterior distribution contains plausible values of parameters. It is then possible to use this distribution to simulate many datasets containing item responses that are plausible observations in the future, termed Posterior Predictive Data (PPD). The PPD is compared back to the original data and the proportion of discrepant PPD data sets is used as an indicator of fit, or misfit. If the model is a good fit to the data, then future data simulated from the model (i.e., PPD) should look very much like observed item responses. Conversely, if the model is a poor fit to the data, then future simulated data will look different from the observed data.

The general procedure for PPMC is as follows. First, PPD data (item responses) are simulated from the parameter posteriors. The data are simulated by randomly drawing parameter values from the posterior distributions and simulating data that would likely arise if these were true parameter values. Another set of parameter values is then drawn randomly and used to generate another data set that would likely arise from this second value set. This process is repeated until the desired number of simulated data sets is generated – usually 1000 times. Next, a comparison is made between the simulated and
observed data. If the data sets are similar, the conclusion is that the model adequately fits
the data (Lynch, 2007).

The comparison between simulated and observed data is often made using a
discrepancy measure. The term *discrepancy measure* is defined as the use of the
discrepancy, or difference, between observed and simulated data in the PPMC analysis
(Meng, 1994). The discrepancy measures must be identified by features of the data
relevant to the type of misfit interested. There is no limit on the number of discrepancy
measures that could be used in PPMC, which illustrates the flexible nature of the
Bayesian method (Lynch, 2007). However, careful consideration should be given to the
choice of discrepancy measure.

Tests using Bayesian $p$-values are available for drawing conclusions regarding the
similarity of the simulated and observed datasets. Let $T(x)$ be a statistic applied to the
observed data, where the observed data is denoted by $x$. The statistic $T( )$ could be any
commonly available measure used in CFA model-data fit (e.g., $\chi^2$ type statistic,
RMSEA), or other descriptors of the data, such as mean or variability of individual
variables. The same statistic is then applied to each of the simulated data sets ($T(x^{sim})$,
where $x^{sim}$ represents the $sim^{th}$ generated data set). This results in one value of the statistic
for the observed data, $T(x)$, and $sim$ values for the simulated data, $T(x^{sim})$, one for each
$x^{sim}$. The Bayesian posterior predictive $p$-value ($ppp$ value) is

$ppp \text{ value} = \text{proportion}(T(x^{sim}) \geq T(x)). \quad (3)$

This $ppp$ value defined in equation (3) is the proportion of simulated data sets
whose function values $T(x^{sim})$ exceed that of the function $T(x)$ applied to the original data.
$ppp$ values close to 0 or 1 indicate model misfit due to the systematic differences between
observed and simulated data. Typically, $ppp$ values less than .05 or greater than .95 are used to flag a misfitting item in the IRT literature (Sinharay, 2006). Ideally, for excellent model-data fit, the $ppp$ value should be around .5, which indicates that 50% of the simulated datasets have discrepancy measures equal to or greater than the value in the original dataset (Muthén & Asparouhov, 2012). If the simulated datasets are similar to the observed dataset, the model is considered a good fit. However, there is no consensus on which decision rule should be used to flag model-data misfit. For instance, Asparouhov (2017) argued that $ppp$ values based on $\chi^2$ statistics greater than .95 should not be considered indicators of poor fit based on their simulations.

To illustrate the approach, consider a PPMC evaluation of global fit of a multidimensional dataset. First, the one-factor model would be fit to the data using MCMC procedures. One thousand posterior predictive data sets would be simulated from the resulting posterior distributions. A $\chi^2$-type statistic would be computed for the observed data, and then for each of the predictive data sets. If the value of the $\chi^2$-type statistic for the predictive data were larger than the value for the observed data 800 out of 1000 times, the $ppp$ value would be $800/1000 = .80$, indicating adequate model-data fit under Muthén and Asparouhov (2010). On the other hand, if the $\chi^2$-type statistic for the predictive data were larger than the value for the observed data only 5 out of 1000 times, the $ppp$ value would be $5/1000 = .005$, indicating poor model-data fit. Besides the choice of the discrepancy statistic, another complication of using the $ppp$ value is that it can be sensitive to small samples (Meng, 1994).

Taken together, there are two general treatments of $ppp$ values: the hypothesis-testing approach (test of model fit) and the use for obtaining diagnostic information for
model fit (Levy, 2011). The hypothesis-testing approach concerns only whether the model fits the data. The latter approach pertains more to understanding which part of the model fits the data and why. For the purpose of model fit evaluation for modeling psychoeducational instruments, use of the ppp as diagnostic information approach is the motivation for using the correlation residual as a discrepancy. This is because such an approach allows for evaluation of the extent to which the model is over- or under-estimating the reproduced bivariate relation. Despite the flexibility on choosing discrepancy measures for ppp, options to evaluate different aspects of model fit is limited in practice, which limits the deeper understanding of estimated models (Levy & Mislevy, 2016; Au & Ames, 2017).

To date, most PPMC for BSEM applications has almost exclusively used a $\chi^2$-type statistic as the discrepancy measure; note that this is only an indicator of overall, not local misfit. Previous studies found that the ppp could be too conservative (de la Horra & Rodriguez-Bernal, 1997; van Kollenburg, Mulder, & Vermunt, 2017). Table 5 presents BCFA applications in psychological or educational instruments: model type, data, and discrepancy measures used for global and local fit evaluation. More specifically, the discrepancy statistic on which the ppp-value is based is included.

In Table 5, note that eleven of the fourteen applications were conducted using Mplus’ BSEM procedure and with the default Mplus prior specification (prior mean of zero and small prior variances). The primary motivation behind most of these applications is to examine factor structure under both the Frequentist ML estimation and the BSEM procedure proposed by Muthén and Asparouhov (2012). Because Mplus does not provide ppp values for local misfit, discussion of local misfit is lacking in those
publications. In fact, the only study that does examine local misfit using correlations as the discrepancy measure was conducted in WinBugs (Levy, 2011). Thus, it is important to the field to illustrate how local fit evaluation can be performed in conjunction with Mplus.

Table 5

Previous BCFA Applications

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Instrument</th>
<th>Data</th>
<th>Statistical Package</th>
<th>$ppp$-Global</th>
<th>$ppp$-Local</th>
</tr>
</thead>
<tbody>
<tr>
<td>de Beer and Bianchi</td>
<td>2017</td>
<td>MBI</td>
<td>7-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Pasha et al.</td>
<td>2017</td>
<td>Survey</td>
<td>5-point Likert</td>
<td>Amos</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Smith et al.</td>
<td>2017</td>
<td>PIH</td>
<td>9-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Gomez</td>
<td>2016</td>
<td>SIAS and SPS</td>
<td>5-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Kurz et al.</td>
<td>2016</td>
<td>BI-AAQ</td>
<td>7-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Barnett et al.</td>
<td>2015</td>
<td>PMSC</td>
<td>4-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Falkenström et al.</td>
<td>2015</td>
<td>WAI-SR</td>
<td>7-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Fong and Ho</td>
<td>2015</td>
<td>UWES-9</td>
<td>7-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Ryoo et al.</td>
<td>2015</td>
<td>TEMA-3</td>
<td>Cog. measure</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Stenling et al.</td>
<td>2015</td>
<td>SMS-II</td>
<td>7-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Blashill et al.</td>
<td>2014</td>
<td>ABCD-SF</td>
<td>Varied</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Fong and Ho</td>
<td>2013</td>
<td>HADS</td>
<td>4-point Likert</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Golay et al.</td>
<td>2012</td>
<td>WISC-IV</td>
<td>Varied</td>
<td>Mplus</td>
<td>$\chi^2$</td>
<td>None</td>
</tr>
<tr>
<td>Ab Hamid et al.</td>
<td>2011</td>
<td>Employee Values</td>
<td>11-point Likert</td>
<td>Amos</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>Levy</td>
<td>2011</td>
<td>IIS</td>
<td>5-point Likert</td>
<td>WinBUGS</td>
<td>SRMR</td>
<td>Correlation</td>
</tr>
</tbody>
</table>
This lack of tools to evaluate fit is problematic. To illustrate why this is so, a more detailed discussion on local fit is provided next.

**Global and Local Fit, the Frequentist Story**

Under the Frequentist framework, a well-established body of literature exists regarding model fit evaluation. The purpose of model fit evaluation is to obtain evidence for, or against, an adequate representation of the data to a specific interpretation of theory. In the confirmatory factor analytic framework, model misspecification may include a relation between a latent factor and a manifest variable that was fixed to zero when it should have been estimated. In such an example, model-data fit evaluation should result in rejection of the specified model. Although conventional Frequentist model-data fit methods can aid in falsification of the proposed model, they cannot provide evidence for discerning the true, population model. Specifically, models that fit equally well will return the same model-data fit results (which is not true for Bayesian framework, as will be discussed later). Therefore, evaluation of fit can only suggest whether the proposed model is a plausible candidate for making inferences and not whether it is the true population model (Hayduk et al., 2007).

Initially, model fit evaluation was tested using the chi-square statistic for exact fit (Jöreskog & Sörbom, 1993). The chi-square test is carried out in the null hypothesis significance testing (NHST) framework, and commonly criticized features of NHST also apply to this test. The main criticism of this approach is that this procedure is a test for perfect fit, which does not align with the goal of finding the most parsimonious model that adequately represents the data. To combat this problem, approximate fit indices were developed as a response that spurred decades of research and discussions. The following
overview will cover chi-square test for exact fit, approximate fit indices, controversy regarding their use and interpretation, and current reporting practice. After a discussion of global fit evaluation, local fit procedure will be presented.

**Chi-square test for exact fit.** A commonly used global fit evaluation approach is the chi-square test for exact fit. The chi-square test for exact fit seeks to compare the proposed model to the theoretically perfect model (Jöreskog & Sörbom, 1993). It is a statistical significance test in which the null hypothesis states that the population covariance matrix of observed variables is not different from the model-implied covariance matrix. However, since the population covariance matrix is not known, the sample observed covariance matrix is used instead. The idea behind the chi-square test is that model misspecification would manifest in the model-implied covariance matrix (matrix containing relations produced using tracing rules with the estimated model parameters). A statistically significant result at a pre-specified Type I error rate could be due to (a) sampling error, resulting in the instability of the sample covariance matrix or (b) model misspecification, which requires further diagnosis of local misfit using the residuals computed as the difference between the observed and model-implied covariance matrix (Hayduk et al., 2007). However, the utility of this chi-square test to evaluate model fit was criticized on both philosophical and statistical grounds.

**Philosophical grounds.** Aside from statistical issues, such as the influence of features of data, the chi-square test of exact fit has been criticized on philosophical grounds. Recall that statistical models are reduction of reality driven by a priori theory of the phenomenon in question. This tautology suggests that regardless of how a model is specified, the data generating mechanism in reality must always be more complex than
the theorized model. As a result, all statistical models must be misspecified to a certain degree, some more misspecified than others. This conception of statistical models is what George Box means when he wrote “all models are wrong” (1976). Here lies the issue with null hypothesis testing. Since it is known that all statistical models must be a reduction of reality, it must be misspecified (e.g., omitted variables, cross-loadings, residual covariance). A test that seeks to answer whether the model differs from a perfect model is thus problematic. Therefore, exactness is no longer emphasized and fit evaluation turns toward approximate.

**Statistical grounds.** Because the chi-square approach uses p value for decision making (rejection or fail to reject the null hypothesis), features of data such as sample size, model size, and non-normality of distributions may influence Type I error rate (Kenny, 2003). A large sample size often results in smaller p values, as it likely detects differences between the observed covariance matrix and the model-implied covariance matrix that are not substantively meaningful. Although a large sample size suggests a stable observed covariance matrix (that the sample covariance matrix is similar to the population covariance matrix), minor model misspecifications (differences between observed and model-implied covariances) would result in a large chi-square value, leading to model rejection. On the other hand, if a study lacks sufficient sample size for a particular model, the lack of statistical significant result may indicate lack of statistical power to detect misspecification (Bentler & Bonnet, 1980). Furthermore, models with a larger number of variables modeled (more falsifiable models) tend to result in larger chi-square statistics (Kenny, 2003; Moshagen, 2012).
Approximate fit. Because the chi-square test for exact fit was criticized on both statistical and philosophical grounds, approximate fit indices were developed to gather further insight into global fit. In contrast to the exact fit approach used in chi-square NHST, approximate fit indices accept that a model will never fit perfectly and seek to quantify the degree of misfit (Kim & Millsap, 2014). Approximate fit indices conceptually seek to compare the proposed model with either (1) a saturated model in which all possible paths in the model are estimated (i.e., absolute fit indices) or (2) a null or independence model in which no correlations between any variables are specified (i.e., incremental fit indices). To get around issues surrounding $p$ value-based decisions, many fit indices were derived using the chi-square statistics or the fit function (a function used to summarize discrepancies between observed covariance matrix and model-implied matrix to assess convergence in estimation). Theoretically, because these indices do not utilize $p$ values, model rejection should not be affected by sample size and other features of data as much as the chi-square test. This notion engendered a whole body of literature in examining the asymptotic behavior of these indices.

Numerous fit indices have been developed since the inception of approximate fit. There is no strict consensus on which indices to report across the discipline and studies; however, the three most popular indices are Root Means Square Error of Approximation (RMSEA), Comparative Fit index (CFI), and Standardized Root Mean Square Residual (SRMR) as recommended by Kline (2011). As such, the following discussion on global fit evaluation will focus on these three indices.

Despite numerous efforts to study various approximate fit indices, not a single study examined the sensitivity of fit indices under the effects of varying degrees of
misspecification, normality, sample size, and estimators until Hu and Bentler (1998). The premise of their study was grounded on the idea that fit indices were developed to detect model misspecification. Therefore, those indices should be sensitive to varying degrees of misspecification. In their study, model misspecification was operationalized into two types: simple and complex. Simple model misspecification is defined as the misspecification at the latent level (relations between latent factors) whereas complex model misspecification concerns the pattern of relations between the indicators (observed variables) with the latent factors. Furthermore, the choice of estimators was also examined (ML, GLS, and ADF). The results suggest that RMSEA, SRMR, and CFI are at least moderately sensitive to model misspecifications among the fit indices examined and less sensitive to estimators, distributions, and sample size. Because these three indices detect different types of misfit, these three indices became a staple in model fit evaluation literature.

Hu and Bentler’s 1998 simulation study found that the RMSEA and CFI are sensitive to the misspecification of factor pattern coefficients (relations between indicators and factors fixed to zero) under different types of model misspecifications and data conditions, whereas the SRMR is most sensitive to misspecification at the structural level (relations between latent factors fixed to zero).

Substantively, RMSEA is based on the noncentrality parameter. The premise is that if the model is correctly specified, the chi-square statistics should equal the degrees of freedom. Therefore, by taking the difference of the observed chi-square statistics and the degrees of freedom of the proposed model, the difference between the two quantities would imply the amount of error accrued due to model misspecification; this is known as
the noncentrality parameter. Therefore, RMSEA indicates the amount of misfit per
degree of freedom given it divides misfit by degrees of freedom.

The CFI rests on the premise that a null model, in which all relations are fixed at
zero, is misspecified. Thus, the noncentrality (difference between the chi-square and the
degrees of freedom) should be large for the null model. Therefore, a proposed model
driven by theory would be less misspecified than the null model so that the noncentrality
should be smaller for the proposed model than the null model. As such, by comparing the
lowering of noncentrality, a greater decrease of misfit would result in a large CFI that
implies adequate fit.

The SRMR, conceptually, is the average correlation residuals (average difference
between the model-implied correlations and the observed correlations). If SRMR is large,
it suggests that some relations between indicators are not adequately reproduced by the
model.

The advantage of using multiple indices is that each value may signal misfit in
different parts of the model that may help theory revision in future studies. Table 6
provides description of the indices and their uses.

Table 6  
**Commonly Reported Approximate Fit Indices**

<table>
<thead>
<tr>
<th>Index</th>
<th>Description</th>
<th>Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSEA</td>
<td>Adjusted fit function value to evaluate error of approximation.</td>
<td>Most sensitive to factor pattern coefficient misspecification.</td>
</tr>
<tr>
<td>SRMR</td>
<td>Conceptually, the average of residual correlations</td>
<td>Most sensitive to latent factor covariance misspecification.</td>
</tr>
<tr>
<td>CFI</td>
<td>Relative improvement of fit over a null or independence model in which no correlations are specified.</td>
<td>Most sensitive to factor pattern coefficient misspecification.</td>
</tr>
</tbody>
</table>
These indices are commonly interpreted and reported in conjunction with the chi-square test for exact fit. Usually, the chi-square test for exact fit is used to signal whether statistically significant misspecifications are detected. Then approximate fit indices are referenced with certain numeric thresholds (similar to the idea of alpha level) to determine global model fit. For the RMSEA, Hu and Bentler (1999) recommended values at or less than .06 are acceptable. For SRMR, values at or less than .08 are considered adequate fit. As for CFI, values at or greater than .95 are considered adequate fit.

For practitioners, it would appear that Hu and Bentler provided evidence supporting certain thresholds to use for model acceptance and rejection. However, Marsh, Hau, and Wen (2004) demonstrated otherwise. In their reanalysis of Hu and Bentler’s 1999 studies, Marsh, Hau, and Wen noticed that many of the models tested in the original study that were classified as model misspecification should in fact be considered correctly specified. This changes the acceptance and rejection rate in their results, and the authors discovered paradoxical behavior of the fit indices overlooked in the original study. In sum, Marsh, Hau, and Wen (2004) argued for a more holistic approach to examine model fit because decision rules (comparing whether observed indices are greater or less than an arbitrary number) do not generalize well across real research situations.

Practitioners should evaluate fit indices beyond comparison with a threshold, and the meaning of these values should be compared against one another to form a better picture of model-data fit. For example, CFI should be examined along with absolute fit indices in relation to whether the observed correlations are low or high, in addition to whether the value exceeds .95. Because the CFI evaluates compares the noncentrality
parameter associated with the proposed model and the null model, CFI can have a terrible value when the observed correlations are low, because the null model can fit well with this type of data condition. However, because absolute fit indices can look great even when correlations are low, if absolute fit indices are only examined, erroneous conclusions might be drawn.

As such, simulation studies regarding these popular and commonly reported indices flourished to understand fit indices’ asymptotic behavior under various conditions. An overview of major studies is presented in Table 7 (the “X” under the fit index indicates the study examined this index). These studies revealed limitations and inconsistent behaviors that further caution the use of generalized numeric cutoff values provided insights into what these indices meant.

Table 7

<table>
<thead>
<tr>
<th>Study</th>
<th>χ²</th>
<th>RMSEA</th>
<th>SRMR</th>
<th>CFI</th>
<th>General Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lai &amp; Green (2016)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>RMSEA and CFI can disagree on model misspecification. Raised issues with interpreting RMSEA and CFI under certain data conditions. For correctly specified models, small degrees of freedom do not result in rejection for large sample size for SRMR and CFI. Supports that for when sample size less than 200, RMSEA should not be reported.</td>
</tr>
<tr>
<td>Taasobshirazi &amp; Wang (2016)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>RMSEA cutoffs should not be trusted under small sample size and small model size. Greater uniqueness variance is associated with smaller fit indices values by varying factor loading sizes and model sizes. RMSEA is sensitive to omitted factor pattern coefficients and clustering of residuals. RMSEA better detects misspecification for small model sizes.</td>
</tr>
<tr>
<td>Kenny (2014)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heene et al. (2011)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Savalei (2012)</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Responses to the rise of fit indices and increasingly complicated results spurred a new discussion in the field: how should a $\chi^2$-type statistic for a test of exact fit and approximate fit indices be weighed in determining model fit? A well-known recommendation came from Paul Barrett (2007). In his argument, models should be evaluated for their empirical predictive utility and less on fit evaluation. However, since the $\chi^2$ test of exact fit is the only statistical test for model evaluation, it should be sufficient to detect model misspecification and that approximate fit indices should be abolished. He denounced approximate fit indices because most indices are derived from the fit functions and $\chi^2$ statistics. Therefore, he believes the $\chi^2$ tests are sufficient to detect model misspecification. Further, he argued that the influence of sample size should not be an issue because it is difficult for applied research to obtain a sample size greater than

Table 7
Studies on Behaviors and Use of Fit Indices – Continued

<table>
<thead>
<tr>
<th>Study</th>
<th>$\chi^2$</th>
<th>RMSEA</th>
<th>CFI</th>
<th>Notes</th>
</tr>
</thead>
</table>
| Chen et al. (2008)            | X       |       |       | RMSEA cutoff and its confidence interval are not generalizable across different models. Sample sizes influences rejection rate. Homogeneity of fit indices are not consistent with varying size of primary and secondary loadings in various model structure common in personality research.
| Beauducel & Wittman (2005)   | X       | X     | x     | Revealed insufficiency in Hu and Bentler (1998)'s simulation study. Sensitivity of different indices to different misspecifications are not generalizable.)
| Fan & Sivo (2005)             | X       | X     | x     | RMSEA improves as more variables are added to the model. CFI showed an opposite pattern. Item parceling improves fit in terms of $\chi^2$, RMSEA, and CFI for unidimensional model.
| Kenny & McCoach (2014)        | X       |       | x     |                                                                       |
| Bandalos (2002)               | X       | X     | x     |                                                                       |
1,000, in his opinion. According to Barrett, should a study contain less than 200 participants, that study should be rejected from publication. However, since Barrett’s claims are bold and lack adequate empirical or simulation studies to support his recommendations, his arguments are called into question, which spawned a great deal of reactions.

Paul Barrett’s vocal criticism of approximate fit indices fueled numerous discussions and publications. Although many commentaries concurred that chi-square test for exact fit has its utility and should not be ignored in model evaluation, outright rejection of approximate fit indices were met with resistance (Markland, 2007; McIntosh, 2007; Millsap, 2007). Many argued that Paul Barrett’s observation of sample size in applied research was short-sighted and unrealistic. Often times, it is possible for applied studies to have over 1,000 participants. Also, many practitioners support the use of approximate fit indices as they provide fit information from different perspectives. Even now, discussion of model fit evaluation is ongoing. For example, interpretation for a significant chi-square test and fit indices under categorical analyses or adjusted chi-square statistics are still contended. Since conventional estimation methods assume multivariate normality and the fit indices assume interval data, instruments that are scored dichotomously (correct/incorrect) can be difficult to study (requiring different estimation methods and different interpretation of fit indices). Therefore, substantive interpretations of fit indices are subject to debate.

**Local fit evaluation.** Local fit is not as frequently discussed as global fit evaluation. Conventionally, if global fit evaluation suggests model misspecification, it is necessary to diagnose which relations are not reproduced well so to determine what part
of the model will need revision to address those misfits (Greiff & Heene, 2017).

Assessment of local fit is often accomplished by examining the modification indices (MI; MacCallum et al, 1992) and/or the covariance residual matrix (Maydeu-Olivares, 2017).

Modification indices estimate the reduction of the chi-square value if a particular path in the model is freely estimated rather than fixed. The modification indices estimate the difference between the chi-square statistics from the proposed model and a model with one more estimated relation. This index is often generated by SEM packages automatically and accompanies the model results. Therefore, less-experienced practitioners may be tempted to adjust their model using these results to improve model-data fit. This approach is dangerous and is not recommended by methodologist for both statistical and substantive reasons. MacCallum et al. (1992) demonstrated that MIs are not stable across samples and capitalize on chance. That is, they are prone to sampling error. Recall that MI estimates the reduction of chi-square statistics using the original chi-square statistics. Therefore, it is possible that MI may pick up on fluctuations due to sampling error rather than model misspecification.

Substantively, adjusting the model based on statistics and not on theory is antithetical to the purpose of the model testing. The purpose of testing a model is to obtain evidence to support a parsimonious explanation of a set of items in an instrument. Fit improvement should be less of a priority, and formulation of solid theory should be of higher concern. If the model is modified with considerations other than theory, results will be difficult, if not possible, to explain. Currently, if MI are used, practitioners must justify the modification on theoretical ground for APA publication (Appelbaum et al., 2018).
Another way to examine local fit is through the covariance residual matrix. The covariance residual matrix is obtained by subtracting the observed covariance matrix and the model-implied covariance matrix. Recall, the model-implied covariances are produced by the set of estimated parameter values taken to be true. If the entire model is correctly specified, the model-implied covariance matrix should be very similar to the observed covariance matrix. The raw residuals can be transformed to a z-score metric to determine statistical significance of misfit. However, the significance test associated with standardized covariance residuals is influenced by sample size, much like null hypothesis significance testing. Therefore, the size of the residuals is often interpreted instead. Ideally, if all observed relations are adequately reproduced, the residuals should be approximately zero. Larger local misfit is accompanied by a larger residual (i.e., |3| or |4|) on z score metric).

Correlation residuals are differences between observed and model-implied correlations. Correlation residuals are often perceived as more interpretable given this simple computation and the familiarity of the correlation metric. Correlation residuals of size less than |.15| are considered non-negligible.

If the overall model is a poor fit to the data, local misfit information can be used to identify where the misfit is originating and possibly provide insight for how to modify the model. For example, if a cluster of items is underestimated and share large inter-item correlations, a separate factor may need to be extracted. It is recommended that standardized discrepancies should be reported to allow external judgement on fit in different areas of model (McDonald & Ho, 2002). Moreover, taken together, global and
local evaluation of fit indicates how well the model represents the data and guides the
decision on the strengths of validity of inferences from the model.

Global and Local Fit, the Bayesian Story

Naturally, models estimated under Bayesian estimation should also be evaluated
at both global and local levels. However, only global fit indices have been implemented
for use in the general audience. Popular statistical software that enables BCFA
estimation, such as Mplus, provides a $p_{pp}$ value based solely on the $\chi^2$ value as the
discrepancy measure. To reiterate, this is only a test of global fit. This discrepancy
statistic is the same as the common global fit measure used in Frequentist SEM
approaches. Recall that the $\chi^2$ value computationally is the fit function multiplied by the
sample size and is used to statistically test for exact fit between model-implied
covariance matrix and the sample covariance matrix (Kline, 2011). For BCFA PPMC, the
value is computed once for the observed data, and again for each of the simulated data
sets. Although $p_{pp}$ value using $\chi^2$ statistics as discrepancy measure is commonly reported
in Frequentist SEM literature, Bayesian inference has little use for the $\chi^2$ value as the sole
purpose of $p$ value is for rejecting null hypotheses. Instead, the value is used in
computation of the $p_{pp}$ value. Because the purpose of PPMC is to compare relevant
features between simulated and observed data, the summarized residual (in the form of
the $\chi^2$ value) is sufficient and has no need to conform to a known statistical distribution
for Frequentist $p$ value.

Examining local misfit is often accomplished by examining correlated residuals in
the model. This approach is described by Muthén and Asparouhov (2012), which many
recent BCFA applications have relied upon for guidance in BCFA analyses. Under this
approach, a series of BSEM models are estimated: a model containing only theoretically relevant paths, a model with cross-loadings, and a model with both cross-loadings and correlated residuals. For the last model, the error covariances between all items are freely estimated (to reflect prior uncertainty in error covariance). Local misfit can then be detected if the error covariance (correlated residual) posterior distributions are “significant,” that is, if an associated highest density interval does not contain zero, this source of possible misfit can be ruled out. Borrowing Gelman et al. (2004)’s framework regarding three methods for Bayesian model fit evaluation (prior sensitivity analysis, posterior predictive model checking, and posterior inferences), modeling for correlated residuals falls under posterior inference at the conceptual level. Consequently, factors that influence posterior distribution of estimates may bias the conclusions made. Thus, interpreting correlated residual estimates for local misfit can be affected by sample size as well as the prior employed (Falkenström et al., 2015). Because Gelman et al. (2004) recommend evaluating fit using multiple approaches, interpreting the posterior inference for correlated residuals may not provide a complete picture of fit. Other methods, such as $ppp$ value that makes use of posterior predictive data, should be explored.

Another way of examining local misfit is with posterior predictive approach. One example of $ppp$ that is not computed from the typical $\chi^2$ value for global fit comes from Levy and Mislevy (2016). In their BCFA model, SRMR and model-based correlations are used to generate $ppp$ values in WinBUG. The model-implied correlation values are computed from the peaks of the posterior distributions of the model. Then, posterior predictive data containing future, plausible item responses are generated using the posterior distributions. The same model is then fit to each posterior predictive dataset to
obtain the model-implied correlation values. The $ppp$ value for model-implied correlations are then computed by examining the proportion of the posterior predictive data that generated model-implied correlation values greater than or equal to the initial SRMR value. These measures are much more desirable than covariance residuals as they directly compare sizes of relation in a correlation metric, which is more interpretable and intuitive. Conceptually, SRMR is the “average” of correlation residuals that summarizes the correlation residuals between the sample and model-implied correlation matrices, with higher value indicating greater misfit. However, by computing the $ppp$ value using correlations, it is more relevant to gauge how well the relation can be reproduced in the simulated data at the local level as it may flagged specific regions in the matrix that signal misfitting. In other words, the $ppp$ value can indicate the proportion of simulated datasets that reproduced the correlation values equal to or greater than the observed correlations. The $ppp$ values are then compared to a threshold (conventionally, $ppp$ values within .05-.95 are not consider misfitting) to determine if the amount of over- or under-prediction suggests model misspecification. Although a promising approach, Levy and Mislevy wrote their own code and this approach is less accessible to applied researchers without technical programming expertise. However, the merit in using correlations to compute $ppp$ values is not to be discounted because it serves as an indicator of strength of relation and is easily interpretable.

**Consequences of Ignoring Misfit**

Ignoring fit evaluation can have serious implications and consequences in both a Frequentist and Bayesian approach. Local misfit can be masked by global fit results (e.g., in Johnston & Finney, 2010). Thus, problematic areas of the theory may go unnoticed for
instrument developers. If these areas are not addressed, the validity of inferences using that model may suffer, regardless of statistical significance of parameter estimates. Recall that models are derived from theory to approximate reality. If the theory-implied model is consistent with the data, parameter estimates can and should be interpreted. On the other hand, if the model is not consistent with the data, as suggested by fit information, the parameter estimates are unable to reproduce the relations observed from the data and thus cannot be used to support a particular theory (Bollen, 1989). This is not to say that model-data consistency is equivalent to model-reality consistency: finding adequate global and local fit index values may not support theory because model-data consistency only refers to the discrepancy between the observed sample covariance matrix and the model-implied covariance matrix. Further, theories may impose unrealistic assumptions that generate models that explain the data well (Godfrey-Smith, 2003). In other words, misfit at either the global or the local level suggests a lack of consistency between the model and data – the pattern of relations implied by the model parameters does not adequately reproduces the data. Because these relations are expressed as parameter estimates, results from badly fitted models are untrustworthy and should not be used for drawing inferences.

Despite the importance of reporting and interpreting local misfit, this aspect of fit evaluation is not widely implemented and accessible to applied researchers under the Bayesian framework. This is particularly troubling for the current explosion of Bayesian application literature in instrument development. Acceptance of faulty models decreases the validity of inference regarding relations between items. As such, there is a need for
accessible methods to assess local misfit. Specifically, I seek to answer the following research question:

RQ1. Do posterior predictive $p$ values, using inter-item correlation residuals as the discrepancy statistics, provide congruent local fit conclusions with Frequentist estimation results for Confirmatory Factor Analysis Model?

RQ2. Which cutoff criteria for posterior predictive $p$ values will lead to the most congruent conclusions between the Frequentist and the Bayesian approach to local fit evaluation?

The present research seeks to explore the efficacy of correlation residual as discrepancy measure via an empirical data sets by comparing congruency of fit conclusions under the Frequentist and the Bayesian approaches. Ideally, local misfit as identified using correlation residuals under the Frequentist approach and $ppp$ values under the Bayesian approach should yield the same conclusion, indicating that the two methods are analogous. As a result, number of flagged item-pairs as misfitting should be equivalent and number of disagreements should be zero.

In the next chapter, I present the methods and data that will be used to evaluate the efficacy of the Bayesian $ppp$ approach.
CHAPTER III

METHODS

The present chapter outlines the participants, data collection procedures, measures, and the methods of analysis for this study. Recall that the current research seeks to evaluate posterior predictive \( p(\text{ppp}) \) values and congruency of local fit conclusions between the Frequentist and the Bayesian approach for confirmatory factor analysis (CFA) models. Specifically, \( ppp \) values computed using correlation residuals will be used to flag misfitting item-pair relations in the Bayesian approach. To provide a more comprehensive perspective, two models will be fitted to the same dataset (a poorly-fitting model and a better-fitting model). The analyses described in this chapter take a two-stage approach: the theoretical factor models will be estimated under Frequentist methods first to obtain conclusions drawn from residual correlations. Then, the data will be fit to the same model in a BCFA framework, and local misfit will be evaluated using the PPMC approach.

As a reminder to the reader, the research questions to be investigated are as follows:

RQ1. Do posterior predictive \( p(\text{ppp}) \) values, using inter-item correlation residuals as the discrepancy statistics, provide congruent local fit conclusions with Frequentist estimation results for Confirmatory Factor Analysis model?

RQ2. Which cutoff criteria for posterior predictive \( p \) values will lead to the most congruent conclusions between the Frequentist and Bayesian approach to local fit evaluation?
This chapter begins by formally presenting information regarding the data used in this study. Then, an overview of Frequentist model evaluation methods and Bayesian model estimation using Markov Chain Monte Carlo (MCMC) will be discussed to compare the misfit conclusions drawn from both approaches. Lastly, because details regarding posterior predictive data and posterior predictive values are the objects of examination, they will be given considerable attention in the latter parts of the chapter.

As Frequentist model results are a precursor to answering RQ1, Frequentist model evaluation and results will be presented prior to a Bayesian model results in Chapter IV.

Participants

Data used in the study were gathered from a mid-Atlantic university’s student assessment day. Assessment Day is a standardized setting in which quality student data are gathered for program assessment, evaluation of university-wide initiatives, and accountability purposes. Furthermore, the tests administered are of a low-stakes nature, as student performance on the tests does not have consequences for the students. Specifically, Fall 2016 first-year student data will be used – all incoming first-year students are required to participate prior to the first day of the semester. For the sample of the Student Opinion Scale (SOS; Fall 2016), only cases with complete data were retained. The final sample size for the SOS is n=773.

Measure

**Student Opinion Scale.** The Student Opinion Scale (SOS) is a 10-item self-report questionnaire administered to students at the end of their entire testing session on Assessment Day (See Appendix A). Each Assessment Day testing session includes more than one assessment. Students report both the effort they invested in their assessments
and their perceived importance of the tasks they completed (Sundre & Thelk, 2007). Specifically, the SOS has two subscales: Importance and Effort.

The SOS is administered to all students who participated in low-stakes assessments on Assessment Day. The data from the SOS scale were administered as a part of the Natural World Test version 9 (NW9) instrument. The NW9 is an 86-item multiple-choice instrument designed to measure student’s quantitative and scientific reasoning skills (NW9; Sundre, Thelk, & Wigtil, 2008; Sundre & Thelk, 2010). The last 20 items of the NW9 were the SOS items. The last 10 items were part of a worry and emotionality scale and were not related to the original SOS items. Thus, only the first 10 SOS items are used in the present study, which are the original SOS items. Descriptive statistics of the Fall 2016 SOS items can be found in Table 8. The Importance and Effort subscales had Cronbach’s alphas of .778 and .840, respectively. These alphas indicate adequate internal consistency within each subscale.

Each of the two subscales (Importance and Effort) contains five items, rated on a 1 to 5 scale. For the SOS, 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree. All items were endorsed, on average, in the moderate to high range of the scale (i.e., mean item scores range from 3.211 to 4.132), as shown in Table 8. Separate scores are calculated for each subscale, with each possible total subscale score ranging from 5 to 25. The mean and standard deviation for the Importance subscale were 17.871 and 3.373. For the Effort subscale, the mean is 19.272 with a standard deviation of 3.454. Sample items include: “I engaged in good effort throughout these tests” and “While taking these tests, I was able to persist to completion of the tasks.” (Sundre & Thelk, 2007, p. 5).
Table 8

Descriptive Statistics for FA16 SOS

<table>
<thead>
<tr>
<th>Items</th>
<th>M</th>
<th>sd</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>3.734</td>
<td>0.881</td>
<td>-0.411</td>
<td>-0.107</td>
</tr>
<tr>
<td>I2</td>
<td>4.132</td>
<td>0.812</td>
<td>-1.062</td>
<td>1.692</td>
</tr>
<tr>
<td>I3</td>
<td>3.581</td>
<td>0.977</td>
<td>-0.671</td>
<td>0.171</td>
</tr>
<tr>
<td>I4</td>
<td>3.449</td>
<td>0.969</td>
<td>-0.442</td>
<td>-0.224</td>
</tr>
<tr>
<td>I5</td>
<td>3.242</td>
<td>0.929</td>
<td>-0.06</td>
<td>-0.157</td>
</tr>
<tr>
<td>I6</td>
<td>4.035</td>
<td>0.873</td>
<td>-0.912</td>
<td>0.826</td>
</tr>
<tr>
<td>I7</td>
<td>3.211</td>
<td>1.036</td>
<td>-0.024</td>
<td>-0.798</td>
</tr>
<tr>
<td>I8</td>
<td>3.865</td>
<td>0.874</td>
<td>-0.797</td>
<td>0.893</td>
</tr>
<tr>
<td>I9</td>
<td>3.867</td>
<td>0.926</td>
<td>-0.794</td>
<td>0.379</td>
</tr>
<tr>
<td>I10</td>
<td>4.027</td>
<td>0.752</td>
<td>-0.925</td>
<td>1.939</td>
</tr>
<tr>
<td>IMP</td>
<td>17.871</td>
<td>3.373</td>
<td>-0.216</td>
<td>0.402</td>
</tr>
<tr>
<td>EFF</td>
<td>19.272</td>
<td>3.454</td>
<td>-0.500</td>
<td>0.239</td>
</tr>
</tbody>
</table>

IMP Cronbach’s alpha = 0.778, EFF Cronbach’s alpha = 0.840

**Procedure**

The following outline sketches the general steps of the methodology that will be used in the current research. Details regarding each step will be expanded upon in subsequent sections.

1. SOS data were evaluated for missing data, multivariate normality, and outliers.
2. The two a priori models (i.e., good-fitting and poorly-fitting) for the instrument were specified and estimated in Mplus using Frequentist estimators. Global and local fit were examined using conventional approaches in the Frequentist framework (e.g., RMSEA, residual correlation matrix).
3. The same models were specified and estimated in Mplus using the Bayes estimator and default priors.
4. Bayesian posterior convergence was assessed, and global fit was examined using the global \textit{ppp} value, provided by default in Mplus.

5. Congruence of the Frequentist and Bayesian parameter values were examined.

6. Posterior distributions for each parameter were saved onto an external file for both SOS models under Bayesian estimation, and the posterior distributions for the completely standardized factor pattern coefficient and latent correlations were retained after removing the burn-in iterations.

7. To ensure the posterior predictive datasets were generated appropriately, data validation was carried out by conducting parameter recovery of item means and standard deviations as well as Frequentist estimates of factor pattern coefficients, error variances, and interfactor correlations.

8. After data validation, 1,000 sets of posterior parameter values were sampled from the posterior distributions to generate posterior predictive data. Then, correlations between items for all posterior predictive datasets were computed.

9. Bayesian posterior predictive correlation residuals were computed by taking the difference between the posterior correlations (values that generated the posterior predictive data) and posterior predictive correlations (values computed from the posterior predictive data).

10. These Bayesian posterior predictive correlation residuals were then compared to the Frequentist correlation residuals (from Step 2) to compute the \textit{ppp} matrix.

11. The Bayesian \textit{ppp} matrices were used to generate heatmaps.

12. Each element in the \textit{ppp} matrices were evaluated under three sets of cutoff criteria. An element coded as “1” indicated misfit whereas “0” indicated fit. The
Frequentist correlation residual matrices were also recoded to “1” and “0” using \(|.15|\) cutoff criteria. The number of misfitting item-pairs flagged was computed by taking the sum for each ppp matrix.

13. The number of disagreements was computed by subtracting the Bayesian and Frequentist matrices from one another (respective to the model) to check for congruency of local fit conclusion.

To clarify the process of posterior predictive model checking for local fit evaluation used in present research, consider the flow chart in Figure 3. To obtain the Bayesian posterior predictive correlation residuals, the correlations obtained in step 4 was subtracted from step 7. Then, these residuals were compared against the Frequentist correlation residuals to compute the ppp values.

Figure 3. Posterior predictive model checking procedure for local fit evaluation.

**Data Analysis**

The first stage of data analysis began with assumption checking. Particularly, multivariate normality and outliers, Mardia’s normalized multivariate kurtosis and Mahalanobis distance were examined. This is a crucial step before model estimation, as features of the data are vital to selecting appropriate estimators. For cognitive measures,
such as the SOS, Mardia’s normalized multivariate kurtosis were used to examine multivariate normality for the data in LISREL 9.4 (Jöreskog & Sörbom, 2015). Mahalanobis distance values were computed to flag possible multivariate outliers using SPSS Version 24. If multivariate normality is not attained, Maximum Likelihood estimator with correction to $\chi^2$ and standard errors will be used. Cases flagged as outliers will be removed.

Both CFA models for the SOS data were then specified and estimated using Mplus (Muthén & Muthén, 1998-2012). Specifically, the following models, driven by a priori theory, were fit to the same dataset.

The first SOS model fitted was a one-factor model, containing only a single motivation construct to explain the SOS items (Figure 4). This model is known to have worse fit than the correlated two-factor model. The SOS follows a correlated two-factor simple structure, as shown in Figure 5. The two correlated latent factors are Importance and Effort, each measured with five items. This model was derived from Sundre and Thelk (2007).

![Figure 4. One-factor structure for the SOS.](image)
Figure 5. Correlated Two-factor structure for the SOS.

**Frequentist Model Evaluation**

This section describes the general procedure in Frequentist model estimation to contextualize the Bayesian approach used to answer the research questions. Prior to conducting Bayesian estimation, the aforementioned factor models using conventional estimation methods will be estimated in Mplus 7.4 (Muthén & Muthén, 1998-2012). Syntax for Frequentist analyses can be found in Appendix B. Syntax for Bayesian analyses can be found in Appendix C.

The Frequentist approach were used to identify item-pairs flagged as misfit with large correlation residuals. That is, items may still correlate meaningfully with one another after controlling for the construct of interest. For example, for the SOS, presence of local misfit could suggest that student responses on the items are not solely dependent on their perceived importance and effort. In other words, the presence of inter-item correlations that are not adequately modeled would suggest that responses to the items are not solely driven by the main construct. In turn, local misfit would manifest upon examination of the residual correlation matrix.

To evaluate global fit under the Frequentist framework, the $\chi^2$ test for exact fit and indices such as the Root Mean Square Error of Approximation (RMSEA), Standardized
Root Mean Square Residual (SRMR), and Comparative Fit Index (CFI) were used. These indices were evaluated with little reliance on conventional decision rules with numerical cutoffs to avoid making dichotomous decisions regarding good or bad fit based on single value. For example, the conventional cutoffs for RMSEA, SRMR, and CFI are .06, .08, and .95 respectively, as reported by Hu and Bentler (1998). For the present study, the decision rules were not used due to the lack of generalizability of conclusions using decision rules (Marsh, Hau, and Wen, 2004). As such, RMSEA, SRMR, and CFI were examined with regards to their sensitivity to different types of misspecifications.

These values were examined for both SOS models prior to the estimation of Bayesian CFAs. Regarding local fit, correlation residuals provided by the Mplus output will be examined for each item-pair. Residual values greater than |.15| indicate local misfit.

Frequentist results of the CFA estimation are found in Tables 9 and 10. SOS items are termed IMP1, IMP2, and so on for the importance items, and EFF1, EFF2, etc., for the effort items. The MLM estimator from Mplus was used to estimate both models. This estimator was chosen due to the lack of multivariate normality suggested by the Mardia’s normalized multivariate kurtosis of 50.305 to correct for $\chi^2$ values and standard errors. Upon examination of Mahalanobis distance, no cases were flagged or removed. The MLM estimator applies the Satorra-Bentler adjustment to obtain standard errors and fit information that are more accurate in the presence of nonnormality. The resulting $\chi^2$ values were 485.161 ($df=35$) for the one-factor model and 204.526 ($df=34$) for the two-factor model.
For examining global fit, the $\chi^2$ tests of exact fit were statistically significant ($p < .001$) but may not be informative due to the large sample size, which typically results in significant $p$ value. The RMSEA was .129 for the one-factor model and .081 for the two-factor model, which may suggest misspecified paths between some items and latent factors for both models. The SRMR, which is conceptually defined as the average size of correlation residuals, was .084 for the one-factor model and .051 for the two-factor model. This indicates that the reproduced relations among items are similar to the observed correlations when the two-factor model was fitted. Lastly, the CFI of .804 for the one-factor model and .926 for the two-factor model suggested that the two-factor model fits relatively better than a null model more so than the one-factor model.

Although the models did not fit, the factor pattern coefficients and standard errors are presented in Tables 9 and 10. All coefficients are statistically significant at $p < .001$. The latent correlation is .682 with a standard error .029 for the two-factor model. Because there was not adequate global model fit based on global fit indices, local fit was examined next to look for evidence of misfit.

Table 9

| One-Factor Model Unstandardized and Standardized Coefficients and Standard Errors |
|----------------------------------|------------------|
|                                 | Coefficient  | Std. Error |
| IMP1                            | 0.606         | (0.688) 0.030 |
|                                 | 0.365         | (0.374) 0.041 |
| IMP2                            | 0.416         | (0.429) 0.039 |
| IMP3                            | 0.493         | (0.532) 0.034 |
| IMP4                            | 0.484         | (0.554) 0.035 |
Table 9
One-Factor Model Unstandardized and Standardized Coefficients and Standard Errors – Continued

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized</th>
<th>Standardized</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>EFF1</td>
<td>0.646</td>
<td>(0.797)</td>
<td>0.029</td>
</tr>
<tr>
<td>EFF2</td>
<td>0.701</td>
<td>(0.804)</td>
<td>0.031</td>
</tr>
<tr>
<td>EFF3</td>
<td>0.618</td>
<td>(0.597)</td>
<td>0.033</td>
</tr>
<tr>
<td>EFF4</td>
<td>0.659</td>
<td>(0.712)</td>
<td>0.031</td>
</tr>
<tr>
<td>EFF5</td>
<td>0.434</td>
<td>(0.578)</td>
<td>0.032</td>
</tr>
</tbody>
</table>

Table 10
Two-Factor Model Unstandardized and Standardized Coefficients and Standard Errors

<table>
<thead>
<tr>
<th></th>
<th>Importance</th>
<th>Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>Std. Error</td>
</tr>
<tr>
<td>IMP1</td>
<td>0.726</td>
<td>(0.825)</td>
</tr>
<tr>
<td>IMP2</td>
<td>0.438</td>
<td>(0.449)</td>
</tr>
<tr>
<td>IMP3</td>
<td>0.581</td>
<td>(0.600)</td>
</tr>
<tr>
<td>IMP4</td>
<td>0.628</td>
<td>(0.677)</td>
</tr>
<tr>
<td>IMP5</td>
<td>0.551</td>
<td>(0.631)</td>
</tr>
<tr>
<td>EFF1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EFF2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EFF3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EFF4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EFF5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A matrices of Frequentist correlation residuals (difference between the observed correlations and model-implied correlations) were created to examine local fit (shown in Appendix D). Residuals that were greater than |.15| were used to suggest relations between items that were inadequately reproduced by the proposed model. When the one-factor model was fitted to the data, six item-pair relations were larger than |.15|. When the two-factor model was fitted, only one item-pair relation was flagged, with the largest local misfit of .157 found between the two items (IMP2 and IMP3) that are negatively worded. This was not surprising, as negatively worded items tend to relate to one another after accounting for the primary latent factor and corroborate with previous model results on this instrument (Myers, 2017). Overall, the correlated two-factor model appears to be a better supported representation of the data consistent with previous literature. Therefore, this model should also demonstrate adequate model-data fit in the Bayesian analysis as well.

**Bayesian Model Estimation, Markov Chain Monte Carlo Approach**

As a reminder about the basic elements of BCFA, a brief overview is provided before moving into the specifics of the study. Unlike Frequentist model estimation that uses Maximum Likelihood-type estimators, Bayesian estimation aligns with Bayesian philosophy in its incorporation of prior information and uncertainty of parameter estimates. To illustrate the Bayesian estimation process, a mountain charting analogy is provided. Bayesian estimation is akin to charting mountain peaks in an unknown landscape with only a map at hand: the sampler being the cartographer, mountain peaks being the most plausible parameter estimate, the map being the posterior distribution based on data and priors, and the unknown landscape being a space of plausible
parameter values (Flaxman, 2010). This is accomplished by using the Markov Chain Monte Carlo (MCMC) algorithm. One such algorithm, the Gibbs sampler, is the default Bayesian estimation method implemented in Mplus 7.4 (Muthén & Muthén, 1998-2012).

This sampler is compatible with Bayesian inferences for models with multiple parameters as it samples values in a joint, multivariate parameter space.

Under Bayesian inference, the goal of estimation is not to obtain a set of unique parameter values but to obtain a posterior distribution for each unknown parameter, representing a range of plausible values. The MCMC sampling method is employed due to its ability to accomplish these three desirable functions: (1) to estimate certain quantities (the posterior mean) of a distribution from random samples, (2) to generate proposal samples from an initial sample, and (3) to generate new samples that are dependent only upon the previous sample (van Ravenzwaaij et al., 2016). This is markedly different from Frequentist’s Maximum Likelihood estimation methods because the ML estimators audition parameter values that maximize the plausibility of observed data instead of the parameter estimate. In other words, the ML-derived point estimates are assumed to be the set that best reproduce the data for a model. In contrast, the Bayesian posterior distribution provides a range of plausible parameter values that mimics the data generating mechanism.

Continuing the landscape charting analogy, under Frequentist estimation methods, the estimation stops when it reaches the peak of the mountain (the maximized likelihood) and that single point estimate is reported. On the other hand, MCMC methods seek to approximate the posterior distribution. The posterior distribution is the set of values
associated with all parts of the unknown landscape, including information at the peak of the mountain as well as the plateaus and mountainsides.

Although the MCMC estimation method is suitable for Bayesian inferences, it is not without its own drawbacks. The downside of this sampler is that if the parameters of interest are correlated, the sampler may become “stuck” in a particular region of the posterior for many iterations (repeatedly proposing the same parameter values). To use the map and mountain peak analogy, the cartographer is trapped somewhere on a plateau on the side of a mountain range with insufficient tools (information) to get out, resulting in a failed attempt to chart the landscape. A second drawback is nonconvergence, which refers to inadmissible parameter estimates that cannot be trusted. Nonconvergence may occur when MCMC is employed because the accepted values are not representative of the posterior distribution: that is, the shape of the simulated values is not similar to the posterior distribution. Metaphorically, the map produced does not look like the actual landscape, missing hills and mountains. This is different from nonconvergence under Frequentist estimation. For ML estimators, nonconvergence occurs if certain parameters are impossible to compute (model identification issues) or if the reproduced covariance matrix is nowhere near the observed covariance matrix, given a criterion. Thus, evaluation of convergence is a crucial and more involved step in Bayesian estimation.

Choice of Priors

For the SOS data set, as recommended by Muthén and Asparouhov (2012), the default, noninformative priors (i.e., $\text{Normal}(0, +\infty)$) were used on the factor pattern coefficients whereas $\text{Inverse-Wishart}(0, -3)$ prior was used for the latent factor covariance to reflect ignorance in the plausible parameter values. Although the Mplus default priors
were used in the present study, no clear justification for the priors was provided in Muthén and Asparouhov (2012). Because part of the purpose of this study is to provide an easily accessible guide to practitioners, Mplus default priors will also be used here, although the need for justification and further investigation is warranted.

**Iterations and Convergence**

For the SOS, two chains were estimated with 100,000 iterations for each parameter. By default, Mplus discards the first half of the chain as burn-ins. As such, the posterior parameter values generated on the output will be constructed based on 50,000 iterations for SOS.

Before evaluating model parameter posteriors, convergence must be evaluated – Mplus offers several ways. Convergence evaluation in Bayesian framework is more nuanced than in Frequentist estimation. In the current literature, it is common to use both statistical and visual methods to evaluate whether convergence is achieved. In the Frequentist framework, convergence is achieved when the residuals between the model-implied covariance matrix and observed covariance matrix are below a certain criterion. In the Bayesian framework, convergence is achieved when the MCMC-generated posterior distribution adequately represents the theoretical shape of the distribution and when the MCMC sampler efficiently produce iterations that only correlate with their immediately preceding iteration (parameter estimates are only correlated with estimates from an iteration before). The present study used the potential scale reduction factor (PSRF; Gelman et al., 2004), trace plots, and autocorrelation plots.

PSRF, as shown in equation (4), is a statistical method to evaluate convergence that provides information on between-chains by within-chain variability, akin to variance
components in ANOVA. This is an appropriate diagnostic tool because multiple series of iterations (chains) were generated simultaneously in this study. Going back to the cartographer analogy, the two chains can be thought of as representing two cartographers charting the same mountain range, and should yield similar results. As such, between-chain variability should be small and not much larger than within-chain variability. As a rule of thumb, a PSRF lower than 1.2 would indicate convergence (Muthén & Asparahou, 2012). For popular statistical package such as Mplus, the PSRF values at each iteration are given to monitor convergence.

$$\text{PSRF} = \sqrt{\frac{W+B}{B}}$$  \hspace{1cm} (4)

However, statistical tools are merely one part of the larger picture in convergence assessment. It is necessary also to examine visual representation of the chains to determine convergence.

Trace plots and autocorrelation plots visually communicate chain characteristics over many iterations. Trace plots show the parameter values from each iteration that the chain has sampled from the landscape. In other words, it is similar to watching the cartographer drawing the map. Ideally, the sampled values should be representative or at least proportional to the target posterior distribution. Therefore, convergence is achieved if the trace plot reveals adequate mixing between chains, as exemplified in Figure 6.
Figure 6. Example of a trace plot that suggests convergence has achieved.
In Figure 7, the chain did not remain in a particular parameter value space for a large number of iterations. This indicates that between each successive iteration, the cartographer is not camping in one area of the mountain for long periods and the chain is not getting “stuck;” rather, it is traversing the parameter space efficiently. Therefore, this chain did not achieve convergence as shown in Figure 7. In addition, if multiple chains for a parameter are specified, the chains should be indistinguishable from one another when superimposed onto the same trace plot (two cartographers should produce similar-looking maps). An autocorrelation plot, on the other hand, seeks to examine the accuracy of the chain. Recall that a desirable property of a Markov chain is its ability to generate successive samples that are only correlated with the immediately previous sample. It
follows that the correlation between samples should sharply decline as the lag increases. Using this method, convergence is achieved if the autocorrelation plot shows a sharp decline as lag increases.

**Foreshadowing BCFA Local Misfit Results**

Recall that the purpose of the present research is to evaluate local misfit using correlation residual to compute posterior predictive $p$ values ($ppp$) under BCFA models. The $ppp$ value indicates the proportion of posterior predictive data (data randomly generated from the posterior distribution) that reproduced the discrepancy statistics (the correlation residuals) equal to or greater than the Frequentist correlation residual values. If the model is a good representation of the observed data, data simulated from the model should be similar to the observed data. A $ppp$ value of .5 would suggest that half of the posterior predictive data reproduced the correlation residual values in the posterior predictive datasets equal to or greater than the Frequentist correlation residual value, which would provide evidence to support the hypothesized model. Ideally, conclusions based on Bayesian local misfit examination should be analogous to conclusions drawn from the Frequentist correlation residual matrix under the same model. That is, if the model is a bad fit to the data, the model should not be able to adequately reproduce the observed bivariate relations under the Frequentist approach, and the same model should not be able to simulate data with similar discrepancy statistics to the observed values under the Bayesian approach. For this study, based on the Frequentist results, local misfit should be obvious in the one-factor model and should not be found in the two-factor model. The same item-pair relations flagged as misfitting in the Frequentist results should also be flagged in the Bayesian results.
Posterior Predictive Data

After the SOS models were estimated using Bayesian estimation, posterior predictive data were simulated. Posterior predictive data are plausible, future data sets that could be observed in the future that fits the model. Recall that Bayesian model estimation provides posterior distributions for each estimated parameter associated with that specific model. 1,000 posterior predictive datasets containing future, plausible item responses if that specified model generated the observed data were generated from 1,000 randomly selected sets of posterior parameter distributions. This process seeks to mimic the data generating mechanism by generating data that aligns with the specified model. To accomplish this procedure, posterior distributions for each parameter were saved to a separate file during model estimation. However, since Mplus does not permit saving of posterior predictive data, the posterior distributions for the factor pattern coefficients, latent factor covariance, and error variances were read into SAS version 9.4. PROC SURVEYSELECT were used to make 1,000 random draws from the posterior. Then, PROC IML will be used to create 1,000 model-implied covariance matrices following the aforementioned CFA models. Using equation (5), the posterior model-implied covariance matrix will be computed.

$$\sum(\hat{\theta}) = \Lambda \Phi \Lambda' + \Theta,$$

(5)

where Lambda ($\Lambda$) is the matrix of factor pattern coefficient, phi ($\Phi$) is the matrix of latent covariances, and theta ($\Theta$) is the matrix of the measurement error variances. Because the posterior distribution file produced by Mplus contains factor pattern coefficient, latent covariances, and error variances, it is possible to use equation (5) to create 1,000 model-implied covariance matrices. In essence, the random draws mimic the
data generating mechanism obtained using prior information and the observed data. Then, the draws were used to create plausible future observations. After the model-implied covariance matrix is computed for all 1,000 draws, Randnormal function in PROC IML were used to simulate multivariate normal posterior predictive data to be used for posterior predictive model checking (PPMC).

**Data Validation**

Before conducting PPMC for local misfit, it is necessary to make sure that the posterior predictive model-implied correlation matrices are generated properly. If the posterior predictive model-implied correlation matrices were properly generated, they should resemble the posterior model-implied correlation matrix from Mplus. However, for the SOS, Mplus outputs only the posterior model-implied covariance matrix. Therefore, to obtain the correlations, the Mplus posterior model-implied covariance matrix will be converted to correlation matrix. A sample of 10,000 data points were simulated using the medians (default point summaries from Mplus) of the posterior distributions of both SOS models. Then, means, factor pattern coefficients, error variances, and interfactor correlation were recovered.

**Discrepancy Statistics**

In the PPMC process, if the CFA model is a good fit to the observed data, then future item response data simulated from the model should look very much like the observed data. Conversely, if the model is a poor fit to the observed data, then future simulated data that align with the model will look different from observed data (Lynch, 2007). Recall that a discrepancy statistic is used to determine how similar the simulated and observed data are (Meng, 1994). It follows that for CFA models, if the factor model
is a good fit to the observed data, Bayesian posterior predictive correlation residuals from the posterior predictive data should be similar to the observed correlation residuals. This is because the posterior predictive data were simulated to fit the model; if the model fits the observed data, then the posterior predictive data will be similar to the observed data, resulting in similar correlation residuals. As such, the correlation residuals between the model and the posterior predictive data should be small. It is the goal of the present research to summarize this information by computing posterior predictive \( p \) values (\( p_{pp} \)) and subsequently present this information using heatmaps.

**Posterior Predictive \( P \) Values**

Using the posterior predictive data, after data validation, posterior predictive correlations will be computed for 1,000 posterior predictive datasets. These posterior predictive correlations will then be subtracted from the posterior correlations (used to generate the posterior predictive data) to obtain the Bayesian posterior predictive residuals. To summarize the correlation residual comparisons, elements in the Bayesian \( p_{pp} \) matrices are coded as “1” if greater than the Frequentist correlation residuals and “0” if it is smaller. The proportions of residuals values that exceed the corresponding Frequentist residuals are obtained by averaging the coded residuals to create the \( p_{pp} \) values. These \( p_{pp} \) values are stored in matrix format and heat maps were then imposed. Fit conclusion matrices were than created by assigning “1” to \( p_{pp} \) values that lies outside the three decision rules: (1) \(.025-.975\), (2) \(.05-.95\), and (3) \(.10-.90\). Values that lie outside these intervals were considered misfitting.
Outcomes

To answer the research questions, two indices were used: the number of item-pair relations flagged as misfitting and the number of disagreements between the Frequentist and the Bayesian fit conclusion matrices. For perfect congruence, the number of item-pair relations flagged as misfitting should be equivalent between the Frequentist and Bayesian results and that the number of disagreements should be zero. The number of disagreements is computed first by subtracting the Frequentist fit conclusion matrix to each Bayesian fit conclusion matrices. Then, the absolute values of the elements were summed to for the number of disagreements. Given that the two-factor model should be a better fit than the one-factor model, the two-factor model Bayesian ppp matrices should flagged fewer item-pairs as misfitting than the one-factor model Bayesian ppp matrices.
CHAPTER IV

RESULTS

The present chapter presents results relevant to the study’s research questions. The research questions are:

(1) Do posterior predictive $p$ ($ppp$) values, using inter-item correlation residuals as the discrepancy statistics, provide congruent local fit conclusions with Frequentist estimation results for Confirmatory Factor Analysis model?

(2) Which cutoff criteria for posterior predictive $p$ values will lead to the most congruent conclusions between the Frequentist and Bayesian approach to local fit evaluation?

The purpose of the present research is to evaluate inter-item correlation residuals as discrepancy measures in detecting local misfit in CFA models. An empirical example using the SOS data is provided in order to illustrate the concept, and because the SOS data fit to CFA models has been extensively examined in previous work (e.g., Myers, 2017, Sundre & Moore, 2002; Thelk et al., 2009). To provide a more comprehensive perspective of local misfit detection, results from a model with good fit to the data (i.e., the two-factor model) will be compared with results from a poorly-fitting model (i.e., the one-factor model) for the SOS data. To gauge the effectiveness of this approach, the congruency of conclusions drawn from the Bayesian approach will be compared with the conclusions drawn from Frequentist approach.

Three numerical cutoff rules for the $ppp$ values are explored in the present study so as to determine which approach is most congruent with the frequentist approach. The three cutoff rules are: (1) .025-.975, (2) .05-.95, and (3) .10-.90. $Ppp$ values outside these
intervals are used to flag local misfit. SAS Syntax to simulate the posterior predictive data and compute $ppp$ values for both SOS models are presented in Appendices E and F, respectively.

This chapter begins by briefly summarizing the methodology before addressing the procedures for generation of the posterior predictive $p$ value matrix for the two SOS models. These procedures involve convergence diagnosis, congruence of Bayesian posterior estimates with Frequentist estimates, and data validation for the posterior predictive data. Then, congruence of fit conclusions between Frequentist approach and posterior predictive $p$ value matrices will be addressed.

**Methodology Outline**

The following methodological procedure was followed for the study:

1. SOS data were evaluated for missing data, multivariate normality, and outliers.
2. The two a priori models (i.e., good-fitting and poorly-fitting) for the instrument were specified and estimated in Mplus using Frequentist estimators. Global and local fit were examined using conventional approaches in the Frequentist framework (e.g., RMSEA, residual correlation matrix).
3. The same models were specified and estimated in Mplus using the Bayes estimator and default priors.
4. Bayesian posterior convergence was assessed, and global fit was examined using the global $ppp$ value, provided by default in Mplus.
5. Congruence of the Frequentist and Bayesian parameter values were examined.
6. Posterior distributions for each parameter were saved onto an external file for both SOS models under Bayesian estimation, and the posterior distributions for
87

the completely standardized factor pattern coefficient and latent correlations were retained after removing the burn-in iterations.

7. To ensure the posterior predictive datasets were generated appropriately, data validation was carried out by conducting parameter recovery of item means and standard deviations as well as Frequentist estimates of factor pattern coefficients, error variances, and interfactor correlations.

8. After data validation, 1,000 sets of posterior parameter values were sampled from the posterior distributions to generate posterior predictive data. Then, correlations between items for all posterior predictive datasets were computed.

9. Bayesian posterior predictive correlation residuals were computed by taking the difference between the posterior correlations (values that generated the posterior predictive data) and posterior predictive correlations (values computed from the posterior predictive data).

10. These Bayesian posterior predictive correlation residuals were then compared to the Frequentist correlation residuals (from Step 2) to compute the ppp matrix.

11. The Bayesian ppp matrices were used to generate heatmaps.

12. Each element in the ppp matrices were evaluated under three sets of cutoff criteria. An element coded as “1” indicated misfit whereas “0” indicated fit. The Frequentist correlation residual matrices were also recoded to “1” and “0” using |.15| cutoff criteria. The number of misfitting item-pairs flagged was computed by taking the sum for each ppp matrix.
13. The number of disagreements was computed by subtracting the Bayesian and Frequentist matrices from one another (respective to the model) to check for congruency of local fit conclusion.

**Model Convergence Diagnosis**

Convergence diagnostic methods such as traceplots, posterior density plots, autocorrelation plots, and potential scale reduction factor (PSRF) values were examined. Traceplots for both SOS models can be found in Figures 8 and 9, showing adequate convergence. Examining the posterior density plot in Figures 10 and 11 shows that all parameter posteriors had only one peak (unimodal distribution) with little abnormalities. Autocorrelations (presented in Figures 12 and 13) tapered off quickly, providing further evidence of convergence. Lastly, the largest PSRF values for both SOS models were equal to one, a commonly accepted heuristic used to indicate convergence of multiple chains. Taking altogether the four pieces of convergence diagnostics, all parameter posteriors, from both models, achieved convergence and further inferences can be made from the posteriors.
Figure 8. Traceplots for the one-factor model.
Figure 9. Traceplots for the two-factor model.
Figure 10. Posterior density plots for the one-factor model.
Figure 11. Posterior density plots for the two-factor model.
Figure 12. Autocorrelation plots for the one-factor model.
Figure 13. Autocorrelation plots for the two-factor model.

**Posterior Parameter Estimate Comparison**

Default, non-informative priors were used to estimate the models under Bayesian framework; little prior knowledge was incorporated into the estimation. As such, Bayesian parameter values should be similar to Frequentist parameter estimates. The parameter value comparison in Figures 14 and 15 compare Frequentist parameter estimates with the Bayesian parameter point-estimates (by default, Mplus reports the median of the posterior distribution). The scatterplots revealed that both parameter
solutions return congruent values as the estimated parameter values were mostly equivalent across the two estimation methods.

Figure 14. One-factor model estimates comparison of the two estimation frameworks.
Data Validation

One thousand posterior predictive data (PPD) sets were simulated from the posterior distributions found in Figures 10 and 11. To ensure that the PPD were generated appropriately, a sample of 10,000 data points were simulated using the medians of the posterior distributions for each SOS model. These validation datasets were then fitted with the corresponding models to recover observed means, factor pattern coefficients, residual variances, and the interfactor correlation (for the two-factor model only).

Figure 15. Two-factor model estimate comparison of the two estimation frameworks.
Successful parameter recovery indicates that the syntax appropriately used the posterior point summaries from the model to simulate data. Figures 16 and 17 contain scatter plots comparing the posterior parameter values to the recovered parameter values. The scatterplots reveal that all parameters were successfully recovered and the PPD were simulated adequately.

*Figure 16. Scatterplot of “true” and recovered parameter values for the one-factor model.*
**Figure 17.** Scatterplot of “true” and recovered parameter values for the two-factor model.

**Visual Comparison of Heatmaps**

Heatmaps of the Bayesian $ppp$ and Frequentist correlation residual values are shown in Figures 18 and 19. The $ppp$ values were computed using equation (3). Values closer to zero are indicated by blue cells and values closer to one are indicated by red cells. Whereas, values closer to .5 are indicated by purple. Frequentist correlation residuals greater than absolute value of .10 were highlighted. From a visual examination of the heatmaps, when the data were fit to the one-factor model, the PPC residuals within subscales were consistently smaller than the corresponding Frequentist correlation.
residuals, as indicated by the ppp values close to zero. The PPC residuals for items across subscales were consistently larger than the Frequentist correlation residuals, as indicated by the ppp values close to one. This general trend is expected since in a one-factor model, only one factor was extracted. Therefore, if the data is multidimensional, common variance shared by that one factor would be small, resulting in smaller within-scale correlations and larger across-scale correlations. This, in turn, would lead to larger discrepancy in opposite directions as shown in the heatmaps.

When the data were fit to the two-factor model, the heatmap for the ppp matrix model followed a similar pattern; the PPC residuals were consistently smaller than the Frequentist counterparts for items within subscales and larger for items across subscales. As expected, the ppp values were more extreme (i.e., closer to zero or one) for the one-factor model heatmap, suggesting that misfit manifested more strongly in the PPD from the one-factor model than PPD from the two-factor model. This is consistent with previous findings in which the two-factor model fits better than then the one-factor model.

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>0.05085</td>
<td>0.16828</td>
<td>0.25758</td>
<td>0.08356</td>
<td>-0.0299</td>
<td>-0.044</td>
<td>-0.0709</td>
<td>-0.0759</td>
<td>-0.0288</td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>0.011</td>
<td>0.26578</td>
<td>-0.0013</td>
<td>0.22434</td>
<td>-0.0501</td>
<td>-0.0874</td>
<td>-0.0423</td>
<td>-0.0243</td>
<td>-0.0066</td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>0.000</td>
<td>0.000</td>
<td>0.1633</td>
<td>0.25105</td>
<td>-0.0781</td>
<td>-0.0894</td>
<td>-0.0964</td>
<td>-0.1001</td>
<td>-0.0532</td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>0.000</td>
<td>0.298</td>
<td>0.000</td>
<td>0.06771</td>
<td>-0.0763</td>
<td>-0.0402</td>
<td>-0.0543</td>
<td>-0.091</td>
<td>-0.0461</td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>0.000</td>
<td>0.000</td>
<td>0.003</td>
<td>-0.055</td>
<td>-0.0318</td>
<td>-0.1163</td>
<td>-0.0868</td>
<td>0.04008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>0.051</td>
<td>0.621</td>
<td>0.821</td>
<td>0.73</td>
<td>0.392</td>
<td>-0.055</td>
<td>0.05306</td>
<td>0.0025</td>
<td>0.03366</td>
<td>0.0344</td>
</tr>
<tr>
<td>EFF2</td>
<td>0.167</td>
<td>0.928</td>
<td>0.919</td>
<td>0.342</td>
<td>0.172</td>
<td>0.000</td>
<td>0.04763</td>
<td>0.03147</td>
<td>-0.0082</td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0.785</td>
<td>0.733</td>
<td>0.98</td>
<td>0.773</td>
<td>0.992</td>
<td>0.009</td>
<td>0.14195</td>
<td>0.0298</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0.719</td>
<td>0.44</td>
<td>0.97</td>
<td>0.93</td>
<td>0.877</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.02739</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>0.207</td>
<td>0.266</td>
<td>0.732</td>
<td>0.559</td>
<td>0.005</td>
<td>0.019</td>
<td>0.332</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 18.* Heatmaps containing ppp and Frequentist correlation residual values for the one-factor model.
To answer Research Questions 1 and 2, two indices were used to quantify the comparison of fit conclusions between Frequentist and Bayesian model results: 1) the number of misfitting item-pairs flagged and 2) the number of disagreements. For the present study, if the Frequentist approach flagged an item-pair relation as misfitting, the same item-pair relation should be flagged as misfitting under the Bayesian approach for an agreement. Therefore, if there is perfect congruence between the Frequentist and Bayesian approach, the item-pairs flagged as exhibiting misfit will be equivalent between the Frequentist correlation residual matrices and the Bayesian \textit{ppp} matrices. With perfect congruence, the number of disagreements will be zero. Departures from exact agreement indicate inconsistencies between the Bayesian and Frequentist approaches.

Three cutoff rules were used to evaluate the \textit{ppp} matrices: (1) .025-.975, (2) .05-.95, and (3) .10-.90. If the \textit{ppp} value fell outside of these intervals, then the region was flagged for local misfit. First, elements within the \textit{ppp} matrices were assigned to either “0” or “1” in which a “1” indicates misfit under one of the three cutoff rules. Similarly,
the Frequentist correlation residual matrices were assigned to “1” or “0” using the |.15|
rule so that residual correlation values less than -.15, or greater than .15, would indicate
misfit. The matrices of “0” and “1” are termed the fit conclusion matrices. These
matrices are presented in Tables 11-13 for when the one-factor model was fitted and
Tables 14-16 for when the two-factor model was fitted. For the number of disagreements,
the fit conclusion matrices between Frequentist and Bayesian framework were subtracted.
The absolute values of each element were then summed to obtain the number of
disagreements.

Table 11
One-Factor Bayesian ppp Fit Conclusion Matrix (.025-.975)

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 12
One-Factor Bayesian ppp Fit Conclusion Matrix (.05-.95)

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 13
One-Factor Bayesian ppp Fit Conclusion Matrix (.10-.90)

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 14
Two-Factor Bayesian ppp Fit Conclusion Matrix (.025-.975)

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 15
Two-Factor Bayesian ppp Fit Conclusion Matrix (.05-.95)

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 16

**Two-Factor Bayesian ppp Fit Conclusion Matrix (.10-.90)**

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-</td>
</tr>
</tbody>
</table>

If the Frequentist and Bayesian methods return similar fit conclusions (i.e., the same item-pairs flagged as misfit or not), the number of flagged item-pairs should be equivalent. As expected, the Two-Factor model resulted in fewer flagged item-pairs than the one-factor model under both approaches. As shown in Table 17, the Bayesian approach resulted in more item-pairs flagged as misfitting than did the Frequentist approach. In other words, most of the PPC residuals for both SOS models were discrepant (over- and under-predicted) from the Frequentist correlation residuals.

Regarding Research Question 2, the ppp rule using .025-.975 cutoff, the most conservative approach, resulted in closest number of flagged items with the Frequentist result. Using this cutoff criterion, 21 item pairs were flagged as misfitting when the one-factor model is fit to the data. In contrast, the Frequentist approach only flagged six item pairs as misfitting. Twenty item pairs were flagged as misfitting under the Bayesian approach when the two-factor model is fit to the data, in contrast to a single item pair flagged under the Frequentist approach.
Next, agreement/disagreements of fit conclusions were examined. Ideally, the number of disagreements should be zero (i.e., with perfect congruence, fit conclusions are the same for all item-pairs) between Frequentist and Bayesian results. The following equation was used to evaluate agreement/disagreement of the matrices:

\[
\text{Congruence}_p = \text{FCF}_{ij} - \text{FCB}_{ij,p}
\]  

(6)

\(\text{FCF}_{ij}\) are elements in the Frequentist Fit Conclusion matrix between items \(i\) and \(j\). \(\text{FCB}_{ij,p}\) are elements in the Bayesian Fit Conclusion matrix between items \(i\) and \(j\) using cutoff \(p\). Since \(\text{FCF}_{ij}\) and \(\text{FCB}_{ij,p}\) can take on values either “0” or “1” (not flagged as misfit and flagged misfit, respectively), taking the difference of the two matrices will result in “0” for agreement and “±1” for disagreement. By taking the sum of the absolute values of \(\text{FCF}_{ij} - \text{FCB}_{ij}\), the result is the number of disagreements between approaches. In Table 18 below, none of the comparisons (with the Frequentist .15 rule) resulted in zero disagreements. Lastly, the .025-.975 cutoff rule, the most conservative approach, resulted in the most-congruent (least disagreement) fit conclusions for both SOS models.

Table 18

<table>
<thead>
<tr>
<th>Disagreement of Fit Conclusion Between Methods</th>
<th>Ppp Rule 1 (.025-.975)</th>
<th>Ppp Rule 2 (.05-.95)</th>
<th>Ppp Rule 3 (.10-.90)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-Factor</td>
<td>15</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Two-Factor</td>
<td>19</td>
<td>20</td>
<td>24</td>
</tr>
</tbody>
</table>
Summary

In sum, the Bayesian $ppp$ approach to detecting local misfit led to a greater number of item-pairs flagged as misfitting than the Frequentist correlation residual approach. Moreover, the .025-.975 cutoff rule led to most congruency of fit conclusions between the paradigms. Put differently, the Bayesian $ppp$ approach did not lead to congruent fit conclusions with the Frequentist approach. Possible explanations to the lack of congruency and significance of results will be discussed thoroughly in Chapter V.
CHAPTER V

DISCUSSION

In Chapter IV, the results provided answers to these research questions:

(1) Do posterior predictive \( p \) \((ppp)\) values, using inter-item correlation residuals as the discrepancy statistics, provide congruent local fit conclusions with Frequentist estimation results for Confirmatory Factor Analysis model?

(2) Which cutoff criteria for posterior predictive \( p \) values will lead to the most congruent conclusions between the Frequentist and Bayesian approach to local fit evaluation?

To answer these research questions, the number of item-pairs flagged as misfitting and the number of disagreements were used. If perfect congruence between Frequentist and Bayesian approach were achieved, the number of item-pairs flagged should be equivalent and the number of disagreements should be zero. For this study, based on the number of flagged item-pairs and the number of disagreements, the Frequentist and Bayesian approaches arrived at very different conclusions about fit, regardless of the cutoff used in the \( ppp \) approach. In this chapter, a brief summary of results and their significance will be presented, followed by limitations and directions for future research.

Summary of Results

Global Fit. Overall, using global fit indices such as \( \chi^2 \) test of exact fit, RMSEA, SRMR, CFI, and the \( ppp \) value using the \( \chi^2 \) statistics revealed inadequate fit at the global level for both the one-factor and the two-factor model under both approaches. Therefore, local fit results were examined to investigate areas of misfit.
**Research Question 1.** Using the Frequentist correlation residual approach, when the one-factor model was fit to the SOS data, six item-pairs were flagged. When the two-factor model was fit to the same data set, one item-pair was flagged. The correlation residuals are used to flag item pairs that suggest the respective model did not adequately reproduce those specific observed correlations. In other words, after accounting for a general motivation factor, or both importance and effort factors, the flagged item pairs remained correlated.

If only the Bayesian *ppp* matrices were examined, it would be difficult to support the two-factor model over the one-factor model because the one-factor model flagged only one additional item pair as exhibiting misfit. The heatmaps were similar under both one-factor and two-factor model (Figures 17 and 18). If only the Frequentist correlation residuals were examined, the difference of the number of item-pairs flagged between the two models was five. As such, there is clear support for the two-factor model over the one-factor model under Frequentist framework. If these results are to be used for adding error covariance terms to the model, the Bayesian *ppp* approach would lead to a much larger model than the Frequentist approach. More specifically, under the Frequentist approach, when the two-factor model is fit to the data, the item pair IMP2 and IMP3 is flagged as misfitting as they share a correlation residual larger than |.15|. To produce a model with better fit, the error covariance could be added to the model to imply the shared covariance not due to two the SOS factors (though, not usually recommended if the addition was not theory-driven).

**Research Question 2.** When using all three *ppp* cutoff rules, the most conservative cutoff rule (.025-.975) led to the most congruence in fit conclusions between
the Frequentist and Bayesian approaches. However, the difference remains large between
the two approaches, even with the most conservative cutoff rule. Moreover, the fit
conclusions between the .025-.975 and .05-.95 rules were very similar.

**Discussion of Findings**

The obtained extreme *ppp* values of the Bayesian approach may be due to the
posterior predictive data (PPD) generation. The generation process responsible for the
posterior predictive data (PPD) did not account for non-normality exhibited in the
observed data. Since the PPD were simulated to be multivariate normal from the posterior
correlation matrices using the Randnormal function in the IML procedure in SAS (refer
to the SAS syntax in Appendices E and F), the PPD would be inherently different from
the observed data for reasons beyond the adequacy of the specified factor structure. The
Randnormal function uses means and covariance matrices to return random draws from a
multivariate normal distribution. To demonstrate this type of discrepancy, Table 19
contains skewness, kurtosis, and Mardia’s normalized multivariate kurtosis value for
every 100th PPD data set. As shown Table 19, although univariate skewness and kurtosis
are similar, Mardia’s normalized multivariate kurtosis is different between the observed
data and the PPDs; the multivariate kurtosis value for the observed data was much larger
(i.e., departing from normality) than the PPD multivariate kurtosis values. Still, non-
normality only impacts chi-square based fit statistics and should not influence parameter
estimate values. Therefore, it is uncertain whether the extreme *ppp* values observed was
due to non-normality.

Because the Mplus Bayes estimator relies upon a normal likelihood, as does the
Frequentist estimator used in the present study, multivariate normal data were simulated.
According to Asparahouv and Muthén (2010), when the $\chi^2$ statistic is used as the discrepancy statistic, PPMC may also detect misspecification due to non-normality. This is a desirable approach according to Asparahouv and Muthén (2010). However, this logic only applies to the $\chi^2$ statistics and further research is required to understand how non-normality impacts the $ppp$ value’s sensitivity to local model misfit evaluation, such as in the present study.
<table>
<thead>
<tr>
<th>Datasets</th>
<th>IMP1 Skewness</th>
<th>IMP1 Kurtosis</th>
<th>IMP2 Skewness</th>
<th>IMP2 Kurtosis</th>
<th>IMP3 Skewness</th>
<th>IMP3 Kurtosis</th>
<th>IMP4 Skewness</th>
<th>IMP4 Kurtosis</th>
<th>IMP5 Skewness</th>
<th>IMP5 Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Data</td>
<td>-0.411</td>
<td>-0.107</td>
<td>-0.671</td>
<td>0.171</td>
<td>-0.442</td>
<td>-0.224</td>
<td>-0.060</td>
<td>-0.157</td>
<td>-0.797</td>
<td>0.893</td>
</tr>
<tr>
<td>1fac_PPD1000</td>
<td>-0.255</td>
<td>-0.339</td>
<td>-0.283</td>
<td>-0.326</td>
<td>-0.208</td>
<td>-0.431</td>
<td>-0.080</td>
<td>-0.214</td>
<td>-0.243</td>
<td>-0.627</td>
</tr>
<tr>
<td>1fac_PPD200</td>
<td>-0.231</td>
<td>-0.367</td>
<td>-0.364</td>
<td>-0.195</td>
<td>-0.077</td>
<td>-0.484</td>
<td>-0.060</td>
<td>-0.205</td>
<td>-0.211</td>
<td>-0.707</td>
</tr>
<tr>
<td>1fac_PPD300</td>
<td>-0.321</td>
<td>-0.322</td>
<td>-0.334</td>
<td>-0.389</td>
<td>-0.275</td>
<td>-0.359</td>
<td>-0.180</td>
<td>-0.408</td>
<td>-0.300</td>
<td>-0.324</td>
</tr>
<tr>
<td>1fac_PPD400</td>
<td>-0.250</td>
<td>-0.477</td>
<td>-0.242</td>
<td>-0.427</td>
<td>-0.233</td>
<td>-0.515</td>
<td>-0.085</td>
<td>-0.505</td>
<td>-0.409</td>
<td>-0.283</td>
</tr>
<tr>
<td>1fac_PPD500</td>
<td>-0.207</td>
<td>-0.509</td>
<td>-0.285</td>
<td>-0.388</td>
<td>-0.093</td>
<td>-0.604</td>
<td>0.001</td>
<td>-0.441</td>
<td>-0.318</td>
<td>-0.358</td>
</tr>
<tr>
<td>1fac_PPD600</td>
<td>-0.351</td>
<td>-0.348</td>
<td>-0.309</td>
<td>-0.355</td>
<td>-0.165</td>
<td>-0.474</td>
<td>-0.004</td>
<td>-0.515</td>
<td>-0.383</td>
<td>-0.460</td>
</tr>
<tr>
<td>1fac_PPD700</td>
<td>-0.249</td>
<td>-0.258</td>
<td>-0.175</td>
<td>-0.578</td>
<td>-0.179</td>
<td>-0.609</td>
<td>-0.017</td>
<td>-0.240</td>
<td>-0.352</td>
<td>-0.425</td>
</tr>
<tr>
<td>1fac_PPD800</td>
<td>-0.443</td>
<td>-0.049</td>
<td>-0.302</td>
<td>-0.426</td>
<td>-0.146</td>
<td>-0.536</td>
<td>-0.160</td>
<td>-0.351</td>
<td>-0.346</td>
<td>-0.244</td>
</tr>
<tr>
<td>1fac_PPD900</td>
<td>-0.331</td>
<td>-0.481</td>
<td>-0.171</td>
<td>-0.614</td>
<td>-0.023</td>
<td>-0.506</td>
<td>0.014</td>
<td>-0.371</td>
<td>-0.376</td>
<td>-0.382</td>
</tr>
<tr>
<td>1fac_PPD1000</td>
<td>-0.174</td>
<td>-0.507</td>
<td>-0.252</td>
<td>-0.483</td>
<td>-0.197</td>
<td>-0.501</td>
<td>-0.084</td>
<td>-0.338</td>
<td>-0.330</td>
<td>-0.532</td>
</tr>
<tr>
<td>2fac_PPD100</td>
<td>-0.126</td>
<td>-0.632</td>
<td>-0.163</td>
<td>-0.444</td>
<td>-0.160</td>
<td>-0.441</td>
<td>-0.041</td>
<td>-0.481</td>
<td>-0.323</td>
<td>-0.432</td>
</tr>
<tr>
<td>2fac_PPD200</td>
<td>-0.262</td>
<td>-0.423</td>
<td>-0.191</td>
<td>-0.374</td>
<td>-0.128</td>
<td>-0.367</td>
<td>-0.114</td>
<td>-0.394</td>
<td>-0.359</td>
<td>-0.383</td>
</tr>
<tr>
<td>2fac_PPD300</td>
<td>-0.196</td>
<td>-0.389</td>
<td>-0.244</td>
<td>-0.396</td>
<td>-0.191</td>
<td>-0.304</td>
<td>-0.038</td>
<td>-0.349</td>
<td>-0.334</td>
<td>-0.198</td>
</tr>
<tr>
<td>2fac_PPD400</td>
<td>-0.332</td>
<td>-0.208</td>
<td>-0.325</td>
<td>-0.413</td>
<td>-0.211</td>
<td>-0.323</td>
<td>-0.160</td>
<td>-0.335</td>
<td>-0.314</td>
<td>-0.642</td>
</tr>
<tr>
<td>2fac_PPD500</td>
<td>-0.216</td>
<td>-0.283</td>
<td>-0.210</td>
<td>-0.581</td>
<td>-0.177</td>
<td>-0.463</td>
<td>-0.099</td>
<td>-0.374</td>
<td>-0.289</td>
<td>-0.510</td>
</tr>
<tr>
<td>2fac_PPD600</td>
<td>-0.300</td>
<td>-0.182</td>
<td>-0.182</td>
<td>-0.468</td>
<td>-0.170</td>
<td>-0.405</td>
<td>-0.067</td>
<td>-0.38</td>
<td>-0.26</td>
<td>-0.332</td>
</tr>
<tr>
<td>2fac_PPD700</td>
<td>-0.197</td>
<td>-0.678</td>
<td>-0.212</td>
<td>-0.462</td>
<td>-0.196</td>
<td>-0.471</td>
<td>-0.110</td>
<td>-0.332</td>
<td>-0.315</td>
<td>-0.611</td>
</tr>
<tr>
<td>2fac_PPD800</td>
<td>-0.297</td>
<td>-0.431</td>
<td>-0.235</td>
<td>-0.466</td>
<td>-0.133</td>
<td>-0.619</td>
<td>-0.146</td>
<td>-0.418</td>
<td>-0.207</td>
<td>-0.637</td>
</tr>
<tr>
<td>2fac_PPD900</td>
<td>-0.321</td>
<td>-0.301</td>
<td>-0.224</td>
<td>-0.573</td>
<td>-0.196</td>
<td>-0.382</td>
<td>-0.159</td>
<td>-0.482</td>
<td>-0.300</td>
<td>-0.424</td>
</tr>
<tr>
<td>2fac_PPD1000</td>
<td>-0.204</td>
<td>-0.417</td>
<td>-0.214</td>
<td>-0.470</td>
<td>-0.189</td>
<td>-0.407</td>
<td>-0.192</td>
<td>-0.170</td>
<td>-0.293</td>
<td>-0.569</td>
</tr>
</tbody>
</table>
Table 19

Skewness, Kurtosis, and Mardia’s Normalized Multivariate Kurtosis of Observed Data and Posterior Predictive Datasets – Continued

<table>
<thead>
<tr>
<th></th>
<th>EFF1 Skewness</th>
<th>EFF1 Kurtosis</th>
<th>EFF2 Skewness</th>
<th>EFF2 Kurtosis</th>
<th>EFF3 Skewness</th>
<th>EFF3 Kurtosis</th>
<th>EFF4 Skewness</th>
<th>EFF4 Kurtosis</th>
<th>EFF5 Skewness</th>
<th>EFF5 Kurtosis</th>
<th>Multivariate Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Data</td>
<td>-1.062</td>
<td>1.691</td>
<td>-0.912</td>
<td>0.826</td>
<td>-0.024</td>
<td>-0.798</td>
<td>-0.794</td>
<td>0.379</td>
<td>-0.925</td>
<td>1.939</td>
<td>50.305</td>
</tr>
<tr>
<td>1fac_PPD100</td>
<td>-0.440</td>
<td>-0.526</td>
<td>-0.475</td>
<td>-0.244</td>
<td>-0.091</td>
<td>-0.476</td>
<td>-0.343</td>
<td>-0.517</td>
<td>-0.229</td>
<td>-0.698</td>
<td>-2.915</td>
</tr>
<tr>
<td>1fac_PPD200</td>
<td>-0.385</td>
<td>-0.533</td>
<td>-0.544</td>
<td>-0.111</td>
<td>-0.091</td>
<td>-0.543</td>
<td>-0.366</td>
<td>-0.356</td>
<td>-0.280</td>
<td>-0.276</td>
<td>-0.701</td>
</tr>
<tr>
<td>1fac_PPD300</td>
<td>-0.583</td>
<td>-0.058</td>
<td>-0.618</td>
<td>0.003</td>
<td>-0.065</td>
<td>-0.666</td>
<td>-0.430</td>
<td>-0.299</td>
<td>-0.421</td>
<td>-0.283</td>
<td>-1.533</td>
</tr>
<tr>
<td>1fac_PPD400</td>
<td>-0.494</td>
<td>-0.367</td>
<td>-0.417</td>
<td>-0.526</td>
<td>-0.125</td>
<td>-0.437</td>
<td>-0.365</td>
<td>-0.510</td>
<td>-0.215</td>
<td>-0.542</td>
<td>-2.758</td>
</tr>
<tr>
<td>1fac_PPD500</td>
<td>-0.473</td>
<td>-0.496</td>
<td>-0.403</td>
<td>-0.397</td>
<td>-0.198</td>
<td>-0.549</td>
<td>-0.406</td>
<td>-0.269</td>
<td>-0.311</td>
<td>-0.404</td>
<td>-2.146</td>
</tr>
<tr>
<td>1fac_PPD600</td>
<td>-0.376</td>
<td>-0.817</td>
<td>-0.462</td>
<td>-0.223</td>
<td>-0.096</td>
<td>-0.371</td>
<td>-0.292</td>
<td>-0.688</td>
<td>-0.234</td>
<td>-0.530</td>
<td>-1.468</td>
</tr>
<tr>
<td>1fac_PPD700</td>
<td>-0.503</td>
<td>-0.407</td>
<td>-0.587</td>
<td>-0.167</td>
<td>-0.126</td>
<td>-0.502</td>
<td>-0.435</td>
<td>-0.334</td>
<td>-0.368</td>
<td>-0.347</td>
<td>-1.202</td>
</tr>
<tr>
<td>1fac_PPD800</td>
<td>-0.516</td>
<td>-0.130</td>
<td>-0.529</td>
<td>-0.050</td>
<td>-0.061</td>
<td>-0.380</td>
<td>-0.404</td>
<td>-0.364</td>
<td>-0.322</td>
<td>-0.074</td>
<td>-0.679</td>
</tr>
<tr>
<td>1fac_PPD900</td>
<td>-0.438</td>
<td>-0.546</td>
<td>-0.504</td>
<td>-0.302</td>
<td>-0.013</td>
<td>-0.445</td>
<td>-0.429</td>
<td>-0.379</td>
<td>-0.305</td>
<td>-0.454</td>
<td>-3.614</td>
</tr>
<tr>
<td>1fac_PPD1000</td>
<td>-0.489</td>
<td>-0.381</td>
<td>-0.384</td>
<td>-0.566</td>
<td>-0.059</td>
<td>-0.501</td>
<td>-0.334</td>
<td>-0.450</td>
<td>-0.201</td>
<td>-0.700</td>
<td>-2.985</td>
</tr>
<tr>
<td>2fac_PPD100</td>
<td>-0.393</td>
<td>-0.653</td>
<td>-0.389</td>
<td>-0.517</td>
<td>-0.078</td>
<td>-0.561</td>
<td>-0.343</td>
<td>-0.455</td>
<td>-0.237</td>
<td>-0.424</td>
<td>-1.871</td>
</tr>
<tr>
<td>2fac_PPD200</td>
<td>-0.442</td>
<td>-0.468</td>
<td>-0.426</td>
<td>-0.287</td>
<td>-0.133</td>
<td>-0.519</td>
<td>-0.380</td>
<td>-0.394</td>
<td>-0.225</td>
<td>-0.497</td>
<td>-1.748</td>
</tr>
<tr>
<td>2fac_PPD300</td>
<td>-0.451</td>
<td>-0.396</td>
<td>-0.443</td>
<td>-0.317</td>
<td>-0.084</td>
<td>-0.582</td>
<td>-0.422</td>
<td>-0.230</td>
<td>-0.359</td>
<td>-0.387</td>
<td>-2.544</td>
</tr>
<tr>
<td>2fac_PPD400</td>
<td>-0.431</td>
<td>-0.633</td>
<td>-0.467</td>
<td>-0.39</td>
<td>-0.062</td>
<td>-0.645</td>
<td>-0.340</td>
<td>-0.510</td>
<td>-0.334</td>
<td>-0.463</td>
<td>-1.434</td>
</tr>
<tr>
<td>2fac_PPD500</td>
<td>-0.418</td>
<td>-0.259</td>
<td>-0.328</td>
<td>-0.591</td>
<td>-0.133</td>
<td>-0.487</td>
<td>-0.357</td>
<td>-0.447</td>
<td>-0.298</td>
<td>-0.212</td>
<td>-2.155</td>
</tr>
<tr>
<td>2fac_PPD600</td>
<td>-0.358</td>
<td>-0.621</td>
<td>-0.431</td>
<td>-0.370</td>
<td>-0.043</td>
<td>-0.635</td>
<td>-0.328</td>
<td>-0.507</td>
<td>-0.334</td>
<td>-0.265</td>
<td>-3.206</td>
</tr>
<tr>
<td>2fac_PPD700</td>
<td>-0.411</td>
<td>-0.614</td>
<td>-0.436</td>
<td>-0.535</td>
<td>-0.180</td>
<td>-0.543</td>
<td>-0.403</td>
<td>-0.391</td>
<td>-0.387</td>
<td>-0.304</td>
<td>-0.835</td>
</tr>
<tr>
<td>2fac_PPD800</td>
<td>-0.456</td>
<td>-0.318</td>
<td>-0.580</td>
<td>-0.126</td>
<td>-0.067</td>
<td>-0.599</td>
<td>-0.454</td>
<td>-0.188</td>
<td>-0.400</td>
<td>-0.073</td>
<td>-2.851</td>
</tr>
<tr>
<td>2fac_PPD900</td>
<td>-0.476</td>
<td>-0.237</td>
<td>-0.553</td>
<td>-0.145</td>
<td>-0.120</td>
<td>-0.582</td>
<td>-0.329</td>
<td>-0.532</td>
<td>-0.302</td>
<td>-0.378</td>
<td>-1.053</td>
</tr>
</tbody>
</table>
| 2fac_PPD1000    | -0.339        | -0.393        | -0.460        | -0.371        | -0.102        | -0.470        | -0.273        | -0.508        | -0.181        | -0.524        | -1.544                

Note, IMP refers to the Importance items of the SOS and EFF refers to Effort items of the SOS.
The present results regarding local fit conclusions using $ppp$ values were not as straightforward as the results from the study from Levy (2011), which is the only study reviewed that employed discrepancy measures to evaluate fit at the local level. In his study, a variety of discrepancy measures were used at both global and local level. Results indicated that although the model recovered certain test statistics adequately (at the global level), model-based correlation $ppp$ values provided additional insights as to which specific variable associations were over- or under-predicted by the model. However, in Levy (2011)’s study, only Bayesian estimation was conducted. Therefore, congruence between Frequentist and Bayesian approach could not be evaluated.

**Significance of Results**

Although the results demonstrated the lack of congruence between the Frequentist and the Bayesian approach, the current research sought to address an issue within the Bayesian framework by proposing and studying the effectiveness of one discrepancy measure (correlation residuals) in a versatile statistical environment (SAS version 9.4; SAS Institute Inc, 2014). The current research demonstrated that it is possible to obtain local fit information using Mplus output. Moreover, evaluation of model-data fit at the local level is crucial in understanding factor structure of psychoeducational instruments. Since a considerable body of literature in this topic was established in the Frequentist framework, it is necessary to provide evidence pertaining to Bayesian approaches to model-data fit. Furthermore, measurement models such as CFA are often the basic building blocks for structural models. Inferences drawn from measurement models must be substantiated before exploring structural relations between constructs. To further
advance Bayesian structural equation efforts, a solid foundation in evaluating
measurement models under Bayesian estimation is mandatory.

**Limitations**

There are several limitations to this study. First, this study used data from only
one instrument and the data are ordinal but treated as continuous. As such, the
effectiveness of *ppp* approach based on this sample may not be generalizable to other
types of data, such as dichotomously-scored instruments, data that are continuous, or data
with different underlying factor structures. However, since previous research was
available for this instrument, it was possible to compare a model with good fit versus a
model with poor fit.

As mentioned earlier, the PPD may be more dissimilar to the observed data not
because of model specification, but because of the PPD data-generation process.
Particularly, non-normality was not accounted for during the PPD generation. Although
Bayesian methods do not invoke asymptotic statistical theories that require stringent
statistical assumptions, certain data characteristics such as non-normality still matter
since Bayesian methods do impose parametric distributions through the priors and
likelihood. In the current study, the PPD were simulated from a normal distribution.

Use of correlation residual as a discrepancy measure may not be appropriate nor
adequate for CFA models. As mentioned throughout, it is vital to use multiple
discrepancy measures to examine various parts of the model. Since the current study only
examined correlation residuals, other forms of fit evaluation (e.g., prior sensitivity
analysis) under Bayesian framework were ignored when forming fit conclusions of the
SOS models. To gain a more comprehensive perspective and substantiated fit
conclusions, it is strongly encouraged that other discrepancy measures or test statistics along with other fit evaluation results be considered when deciding whether the model fits the data. Moreover, the interpretation of \( ppp \) values can be difficult since the present study did not compare the absolute values of the correlation residuals. Thus, the directions of the residuals made the interpretations of model misspecification difficult.

**Directions for Future Research**

A few suggestions are provided here to further local fit discussion in Bayesian framework. There is no doubt a need to expand accessible options for evaluating model-data fit in structural equation modeling work. Local fit is often neglected in BCFA literature while popular statistical packages such as Mplus (Muthén & Muthén, 1998-2012) provide only global fit information by default. For the present research, obtaining local fit \( ppp \) values requires considerable programming efforts, which may not be feasible for practitioners. Moving forward, other accessible platforms such as R and Stan (Stan Development Team, 2017) should be extended to provide easy-to-use packages to increase local fit results reporting.

There is also a need to explore the efficacy of the PPMC approach in detecting various forms of model misspecifications using multiple discrepancy measures. Drawing from the present results, it is suspected that non-normality may play a critical role in extreme \( ppp \) values. Whether these conclusions about misfit are due to model misspecification or issues with PPD simulation should be explored further in comprehensive simulation studies.
Conclusion

In sum, the *ppp* approach to local fit evaluation used in this research did not provide similar fit conclusions with Frequentist results based on the number of flagged item-pairs and the number of disagreements. Nonetheless, the current project sought to address an area within Bayesian applications that is necessary for Bayesian methods to be widely adopted properly for valid inferences to be drawn. By examining fit at the local level, a greater understanding of psychoeducational instruments can be obtained, regarding relations among latent constructs and observed item responses. Although the approach in this study was not congruent with the Frequentist results, the current methodology sets up future research for more rigorous examination of PPMC approach for measurement models.
Appendix A

Student Opinion Scale

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below.

1. Doing well on these tests was important to me.
2. I engaged in good effort throughout these tests.
3. I am not curious about how I did on these tests relative to others.
4. I am not concerned about the scores I receive on these tests.
5. These were important tests to me.
6. I gave my best effort on these tests.
7. While taking these tests, I could have worked harder on them.
8. I would like to know how well I did on these tests.
9. I did not give these tests my full attention while completing them.
10. While taking these tests, I was able to persist to completion of the tasks.
Appendix B

MPLUS Syntax for the SOS Frequentist

title: SOS 2-factor (mlm);
  data: file is sos2cleanay3.dat;
  variable:
    names are IMP1 IMP2 IMP3 IMP4 IMP5 EFF1 EFF2 EFF3 EFF4 EFF5;

  analysis:
    estimator=mlm;

  model:
    Imp by IMP1* IMP2 IMP3 IMP4 IMP5 ;
    Eff by EFF1* EFF2 EFF3 EFF4 EFF5 ;
    Imp with Eff;
    IMP@1;
    EFF@1;

  output:
    standardized;
    residual;
Appendix C

MPLUS Syntax for the SOS Bayesian

title: SOS 2-factor (Bayes);
data: file is sos2cleanay3.dat;
variable:
names are IMP1 IMP2 IMP3 IMP4 IMP5 EFF1 EFF2 EFF3 EFF4 EFF5;

analysis:
estimator=BAYES;
PROC = 2;
Chains = 2;
FBITER = 100000;

model:
Imp by IMP1* IMP2 IMP3 IMP4 IMP5 ;
Eff by EFF1* EFF2 EFF3 EFF4 EFF5 ;
Imp with Eff;
IMP@1;
EFF@1;

output:
stdyx;
residual;
tech4 tech5 tech8;

SAVEDATA:
BPARAMETERS = SOS_Posterior 9(free).dat;
format is free;

PLOT:
type = plot2;
## Appendix D

### Frequentist Correlation Residual Matrices

#### One-Factor Model Frequentist Correlation Residual Matrix

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>0.051</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>0.168</td>
<td>0.266</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>0.258</td>
<td>-0.001</td>
<td>0.163</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>0.084</td>
<td>0.224</td>
<td>0.251</td>
<td>0.068</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>-0.030</td>
<td>-0.050</td>
<td>-0.078</td>
<td>-0.076</td>
<td>-0.055</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>-0.044</td>
<td>-0.087</td>
<td>-0.089</td>
<td>-0.040</td>
<td>-0.032</td>
<td>0.053</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>-0.071</td>
<td>-0.042</td>
<td>-0.096</td>
<td>-0.054</td>
<td>-0.116</td>
<td>0.003</td>
<td>0.048</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>-0.076</td>
<td>-0.024</td>
<td>-0.100</td>
<td>-0.091</td>
<td>-0.087</td>
<td>0.034</td>
<td>0.031</td>
<td>0.142</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>-0.029</td>
<td>-0.007</td>
<td>-0.053</td>
<td>-0.046</td>
<td>0.040</td>
<td>0.034</td>
<td>-0.008</td>
<td>-0.030</td>
<td>0.027</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Two-Factor Model Frequentist Correlation Residual Matrix

<table>
<thead>
<tr>
<th></th>
<th>IMP1</th>
<th>IMP2</th>
<th>IMP3</th>
<th>IMP4</th>
<th>IMP5</th>
<th>EFF1</th>
<th>EFF2</th>
<th>EFF3</th>
<th>EFF4</th>
<th>EFF5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMP1</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP2</td>
<td>-0.062</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP3</td>
<td>-0.032</td>
<td>0.157</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP4</td>
<td>0.065</td>
<td>-0.106</td>
<td>-0.015</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMP5</td>
<td>-0.056</td>
<td>0.148</td>
<td>0.110</td>
<td>-0.065</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF1</td>
<td>0.056</td>
<td>-0.004</td>
<td>-0.073</td>
<td>-0.032</td>
<td>0.033</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF2</td>
<td>0.042</td>
<td>-0.041</td>
<td>-0.085</td>
<td>0.004</td>
<td>0.056</td>
<td>0.011</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF3</td>
<td>-0.014</td>
<td>-0.011</td>
<td>-0.097</td>
<td>-0.027</td>
<td>-0.056</td>
<td>-0.038</td>
<td>0.006</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EFF4</td>
<td>-0.006</td>
<td>0.014</td>
<td>-0.100</td>
<td>-0.057</td>
<td>-0.013</td>
<td>-0.012</td>
<td>-0.016</td>
<td>0.099</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>EFF5</td>
<td>0.044</td>
<td>0.033</td>
<td>-0.041</td>
<td>-0.005</td>
<td>0.112</td>
<td>0.021</td>
<td>-0.023</td>
<td>-0.047</td>
<td>0.008</td>
<td>-</td>
</tr>
</tbody>
</table>
Appendix E

Two-Factor Model: Posterior Predictive Data and \textit{ppp} Syntax

/\texttt{*observed means*/}
data obsmeans;
input m1 m2 m3 m4 m5 m6 m7 m8 m9 m10;
cards;
run;
/\texttt{*observed sd*/}
data std;
input sd1 sd2 sd3 sd4 sd5 sd6 sd7 sd8 sd9 sd10;
cards;
0.881 0.977 0.969 0.929 0.874 0.812 0.873 1.036 0.926 0.752;
run;
/\texttt{*reading in 1000 posterior draws of all parameters from mplus*/}
proc import datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAPILES\Chi\Thesis\posterior\sos1facposdraw (all).csv"
out=sospost
dbms=csv replace;
getnames=yes;
run;
/\texttt{*creating dataset with just completely standardized loadings with row number*/}
data load;
set sospost;
keep row V42 V43 V44 V45 V46 V47 V48 V49 V50 V51;
row=_n_;

/*create residual variance dataset by taking the R2*/
data err;
set sospost;
keep row V52 V53 V54 V55 V56 V57 V58 V59 V60 V61;
row=_n_;

%macro create(n);
%do i=1 %to &n;
data load_sim;
set load (where=(row=&i));
drop row;

data err_sim;
set err (where=(row=&i));
drop row;

proc iml;
use load_sim;
read all var _NUM_ into load[c=VarNames];
close load_sim;

use err_sim;
read all var _NUM_ into err[c=VarNames];
close err_sim;
use obsmeans;
read all var _NUM_ into om[c=VarNames];
close obsmeans;

use std;
read all var _NUM_ into sd[c=VarNames];
close std;

/*Get the model-implied cov matrix*/
S=I(1); /*interfactor correlation as 1s*/
/*ev=I(10)-diag(err);*/
ev=diag(err); /*create error covariance matrix with just errs on the diagonal*/
sigma=load`*S*(load)+ev; /*the model-implied correlation matrix*/
Jmeans=J(1,10,0); /*creating a 1X10 matrix of just 0s because the mean of z distribution is zero*/
/*Output will be z-scores*/
Xz = randnormal(773, Jmeans, sigma); /*simulate 773 cases (z-scores because everything was standardized) using means of zero and from the model-implied correlation matrix*/
/*Convert back to scale of original data*/
tau=repeat(om, 773, 1); /*create 773X1 matrix of the observed means*/
X=round(tau+Xz#sd); /*unstandardized the simulated z scores using observed means and standard deviations, # sign is cell to cell multiplication*/
create MBS from sigma;
append from sigma;
close MBS;

create MyData from X; /** create data set **/
append from X; /** write data in vectors **/
close MyData; /** close the data set **/
quit;
data MBS_sim&i; set MBS; run;
Data mydata; set mydata;
array col[10];
do j=1 to 10;
if col[j] GE 5 then col[j]=5; /*to prevent out of range responses*/
if col[j] LE 1 then col[j]=1;
col[j]=round(col[j],1); /*round to whole number*/
end;
drop j;
run;
/*creating the posterior predictive correlations*/
proc corr data=mydata outp=corrs&i noprob nosp1 noprint;
var COL1-COL10;
run;
%end;
%mend create;
%create(1000)
/*proc datasets library=work; */
/*delete mydata:; run;*/
/*stacking the corr matrices*/
data S_all;
set MBS_sim:;
rename COL1=MBS1;
rename COL2=MBS2;
rename COL3=MBS3;
rename COL4=MBS4;
rename COL5=MBS5;
rename COL6=MBS6;
rename COL7=MBS7;
rename COL8=MBS8;
rename COL9=MBS9;
rename COL10=MBS10;
run;

data corr_all;
set corrs:; run;
/*only retain relevant rows and create ids for each column*/
data corr_all; set corr_all(where=(_TYPE_='CORR'));
id=1;
if _NAME_='COL2' then id=2;
if _NAME_='COL3' then id=3;
if _NAME_='COL4' then id=4;
if _NAME_='COL5' then id=5;
if _NAME_='COL6' then id=6;
if _NAME_='COL7' then id=7;
if _NAME_='COL8' then id=8;
if _NAME_='COL9' then id=9;
if _NAME_='COL10' then id=10;
run;

proc sort data=corr_all; by id;
data corr_all;
set corr_all;
by id;
if first.id then draw=0;
draw+1;
run;

proc sort data=corr_all; by draw id; run;
/*putting together posterior correlation matrix S_all with the PPD correlations corr_all*/
data PPD;
merge S_all corr_all; run;
/*reading in freq correlation residuals*/
data FR;
input id imp1 imp2 imp3 imp4 imp5 eff1 eff2 eff3 eff4 eff5;
cards;
1 0 0.050848 0.168278 0.257584 0.083558 -0.029856
   -0.044022 -0.070906 -0.075906 -0.028834
2 0.050848 0 0.265784 -0.001308 0.224344 -0.050068
   -0.087436 -0.042338 -0.024318 -0.006582
```plaintext
proc sort data=PPD; by id;
data PPD2;
merge FR PPD;
by id;
proc sort data=PPD2;
by draw id; run;
data PPD2; set PPD2;
br1=COL1-MBS1;
br2=COL2-MBS2;
br3=COL3-MBS3;
br4=COL4-MBS4;
br5=COL5-MBS5;
br6=COL6-MBS6;
br7=COL7-MBS7;
br8=COL8-MBS8;
br9=COL9-MBS9;
br10=COL10-MBS10;
run;
/*making PPP matrix*/
data PPD3; set PPD2; /*If Obs > MI then r=1*/
array ppc_r[10];
do i=1 to 10;
ppc_r[i]=0;
end;
if br1 GE imp1 then ppc_r1=1;
if br2 GE imp2 then ppc_r2=1;
if br3 GE imp3 then ppc_r3=1;
if br4 GE imp4 then ppc_r4=1;
if br5 GE imp5 then ppc_r5=1;
if bre6 GE eff1 then ppc_r6=1;
if bre7 GE eff2 then ppc_r7=1;
if bre8 GE eff3 then ppc_r8=1;
if bre9 GE eff4 then ppc_r9=1;
if bre10 GE eff5 then ppc_r10=1; run;
proc print data=PPD3 (obs=20); run;
```

```
<table>
<thead>
<tr>
<th></th>
<th>0.168278</th>
<th>0.265784</th>
<th>0</th>
<th>0.163302</th>
<th>0.251054</th>
<th>-0.078103</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.089376</td>
<td>-0.096373</td>
<td>-0.100058</td>
<td>-0.053202</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.257584</td>
<td>-0.001308</td>
<td>0.163302</td>
<td>0</td>
<td>0.067712</td>
<td>-0.076284</td>
</tr>
<tr>
<td></td>
<td>-0.040218</td>
<td>-0.054304</td>
<td>-0.091044</td>
<td>-0.046066</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.083558</td>
<td>0.224344</td>
<td>0.251054</td>
<td>0.067712</td>
<td>0</td>
<td>-0.055018</td>
</tr>
<tr>
<td></td>
<td>-0.031796</td>
<td>-0.116308</td>
<td>-0.086838</td>
<td>0.040078</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>-0.029856</td>
<td>-0.050068</td>
<td>-0.078103</td>
<td>-0.076284</td>
<td>-0.055018</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.053062</td>
<td>0.002501</td>
<td>0.033656</td>
<td>0.034404</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.044022</td>
<td>-0.087436</td>
<td>-0.089376</td>
<td>-0.040218</td>
<td>-0.031796</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.053062</td>
<td>0</td>
<td>0.047632</td>
<td>0.031472</td>
<td>-0.008202</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.070906</td>
<td>-0.042338</td>
<td>-0.096373</td>
<td>-0.054304</td>
<td>-0.116308</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.002501</td>
<td>0.047632</td>
<td>0</td>
<td>0.141946</td>
<td>-0.029826</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.075906</td>
<td>-0.024318</td>
<td>-0.100058</td>
<td>-0.091044</td>
<td>-0.086838</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.033656</td>
<td>0.031472</td>
<td>0.141946</td>
<td>0</td>
<td>0.027394</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.028834</td>
<td>-0.006582</td>
<td>-0.053202</td>
<td>-0.046066</td>
<td>0.040078</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.034404</td>
<td>-0.008202</td>
<td>-0.029826</td>
<td>0.027394</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

```plaintext
proc means data=PPD3 mean;
class id;
var ppc_r;
output out=ppc_out;
run;

data ppc_out2; set ppc_out (where=(_STAT_="MEAN" AND _TYPE_=1));
run;
```
Appendix F
One-Factor Model: Posterior Predictive Data and ppp Syntax

/*observed means*/
data obsmeans;
input m1 m2 m3 m4 m5 m6 m7 m8 m9 m10;
cards;
rn;
/*observed sd*/
data std;
input sd1 sd2 sd3 sd4 sd5 sd6 sd7 sd8 sd9 sd10;
cards;
0.881 0.977 0.969 0.929 0.874 0.812 0.873 1.036 0.926 0.752;
rn;
/*reading in 1000 posterior draws of all parameters from mplus*/
proc import datafile="N:\AA\CARS\CARS-Common\Graduate Students\GAPFILES\Chi\Thesis\posterior\sos2facposdraws (all).csv"
out=sospost
dbms=csv replace;
getnames=yes;
rn;
/*creating dataset with just completely standardized loadings with row number*/
data load; set sospost;
keep row V43 V44 V45 V46 V47 V48 V49 V50 V51 V52;
row=_n_; /*creating dataset duplicating the load dataset on top of each other*/
data load2; set load load;
proc sort data=load2; by row;
/*creating loading matrix from each row*/
data load2; set load2;
by row;
if first.row then do;
V48=0; V49=0; V50=0; V51=0; V52=0;
end;
else do;
V43=0; V44=0; V45=0; V46=0; V47=0;
end; run;
/*create residual variance dataset by taking the R2*/
data err; set sospost;
keep row V53 V54 V55 V56 V57 V58 V59 V60 V61 V62;
row=_n_; /*interfactor correlation dataset*/
data corr; set sospost;
keep row V64;
row=_n_; %macro create(n);
%do i=1 %to &n;
data load_sim;
set load2 (where=(row=&i));
drop row;

data err_sim;
  set err (where=(row=&i));
  drop row;

data corr_sim;
  set corr (where=(row=&i));
  drop row;
run;

proc iml;
use load_sim;
read all var _NUM_ into load[c=VarNames];
close load_sim;

use err_sim;
read all var _NUM_ into err[c=VarNames];
close err_sim;

use corr_sim;
read all var _NUM_ into c;
close corr_sim;

use obsmeans;
read all var _NUM_ into om[c=VarNames];
close obsmeans;

use std;
read all var _NUM_ into sd[c=VarNames];
close std;

/*Get the model-implied cov matrix*/
JJ=J(2,1,1); /*creating a 2X1 matrix called "JJ" with just 1s*/
S=JJ*c; /*multiply JJ with the interfactor correlations*/
S=S||S; /*concatenate both matrices*/
S[1,1]=1; /*fixing diagonals to 1*/
S[2,2]=1; /*fixing diagonals to 1*/
/*ev=I(10)-diag(err);*/
ev=diag(err); /*create error covariance matrix with just errs on the diagonal*/
sigma=load`*S*(load)+ev; /*the model-implied correlation matrix*/
Jmeans=J(1,10,0); /*creating a 1X10 matrix of just 0s*/

/*Output will be z-scores*/
Xz = randnormal(773, Jmeans, sigma); /*simulate 773 cases (z-scores because everything was standardized) using means of zero and from the model-implied correlation matrix*/

/*Convert back to scale of original data*/
tau=repeat(om,773,1); /*create 773X1 matrix of the observed means*/
X=round(tau+Xz#sd); /*unstandardized the simulated z scores using observed means and standard deviations, # sign is cell to cell multiplication*/
create MBS from sigma;
append from sigma;
close MBS;

create MyData from X; /** create data set **/
append from X;       /** write data in vectors **/
close MyData;       /** close the data set **/
quit;

data MBS_sim&i; set MBS; run;

Data mydata; set mydata;
array col[10];
do j=1 to 10;
if col[j] GE 5 then col[j]=5; /*to prevent out of range responses*/
if col[j] LE 1 then col[j]=1;
col[j]=round(col[j],1); /*round to whole number*/
end;
drop j;
run;
/*creating the posterior predictive correlations*/
proc corr data=mydata outp=corrs&i noprob nosimple noprint;
var COL1-COL10;
run;

%end;
%mend create;

%create(1000)

/*proc datasets library=work; */
/*delete mydata; run;*/
/*stacking the corr matrices*/
data S_all;
set MBS_sim;..
rename COL1=MBS1;
rename COL2=MBS2;
rename COL3=MBS3;
rename COL4=MBS4;
rename COL5=MBS5;
rename COL6=MBS6;
rename COL7=MBS7;
rename COL8=MBS8;
rename COL9=MBS9;
rename COL10=MBS10;
run;

data corr_all;
set corrs; run;
/*only retain relevant rows and create ids for each column*/
data corr_all; set corr_all(where=(_TYPE_='CORR'));
  id=1;
  if _NAME_='COL2' then id=2;
  if _NAME_='COL3' then id=3;
  if _NAME_='COL4' then id=4;
  if _NAME_='COL5' then id=5;
  if _NAME_='COL6' then id=6;
if _NAME_='COL7' then id=7;
if _NAME_='COL8' then id=8;
if _NAME_='COL9' then id=9;
if _NAME_='COL10' then id=10;
run;

proc sort data=corr_all; by id;
data corr_all;
set corr_all;
by id;
if first.id then draw=0;
draw+1;
run;

proc sort data=corr_all; by draw id; run;
/*putting together posterior correlation matrix S_all with the PPD correlations corr_all*/
data PPD;
merge S_all corr_all; run;
/*reading in freq correlation residuals*/
data FR;
input id imp1 imp2 imp3 imp4 imp5 eff1 eff2 eff3 eff4 eff5;
cards;
1 0 -0.062265 -0.03157 0.065075 -0.055865 0.0559817 0.04156785 -0.0135142 -0.0057869 0.04418095
2 -0.062265 0 0.15683 -0.106313 0.148221 -0.003701196 -0.041207158 -0.11012 0.11012 -0.0725524 -0.0845052
3 -0.03157 0.15683 0 -0.01467 0.013531372 0.032902214 -0.0413484 0.010768 -0.037906 0 -0.0413484
4 0.065075 -0.106313 -0.01467 0 -0.064747 -0.031808908 0.003825666 -0.026656392 -0.056698644 -0.04978978
5 -0.055865 0.148221 0.11012 -0.064747 0 0.032778876 0.056005798 -0.055824776 -0.031425132 0.111982666
6 0.0559817 -0.003701196 -0.0725524 -0.031808908 0.032778876 0 0.010768 -0.037906 -0.012092 0.020776
7 0.04156785 -0.041207158 -0.0845052 0.003825666 0.056005798 0.010768 0 0.005752 -0.016006 -0.022977
8 -0.0135142 -0.011364904 -0.0972376 -0.026656392 0.055824776 -0.037906 0.098522 0.003825666 -0.047116
9 -0.0057869 0.013531372 -0.0998732 -0.056698644 0.013425132 -0.012092 -0.098522 0.008488 0.020776
10 0.04418095 0.032902214 -0.0413484 -0.004978978 0.111982666 0.020776 -0.022977 -0.047116 0.008488
; run;

proc sort data=PPD; by id;
data PPD2;
merge FR PPD;
by id;
proc sort data=PPD2; by draw id; run;
data PPD2; set PPD2;
bril=COL1-MBS1;
bri2=COL2-MBS2;
bri3=COL3-MBS3;
br14=COL4-MBS4;
br15=COL5-MBS5;
br16=COL6-MBS6;
br17=COL7-MBS7;
br18=COL8-MBS8;
br19=COL9-MBS9;
br20=COL10-MBS10;

run;
/*making PPP matrix*/
data PPD3; set PPD2; /*If Obs > MI then r=1*/
array ppc_r[10];
do i=1 to 10;
ppc_r[i]=0;
end;
if bri1 GE imp1 then ppc_r1=1;
if bri2 GE imp2 then ppc_r2=1;
if bri3 GE imp3 then ppc_r3=1;
if bri4 GE imp4 then ppc_r4=1;
if bri5 GE imp5 then ppc_r5=1;
if bre1 GE eff1 then ppc_r6=1;
if bre2 GE eff2 then ppc_r7=1;
if bre3 GE eff3 then ppc_r8=1;
if bre4 GE eff4 then ppc_r9=1;
if bre5 GE eff5 then ppc_r10=1; run;

proc print data=PPD3 (obs=20); run;

proc means data=PPD3 mean;
class id;
var ppc_r;
output out=ppc_out;
run;

data ppc_out2; set ppc_out (where=(_STAT_="MEAN" AND _TYPE_=1));
run;
References


Hamra, G., Richardson, D., MacLehose, R., & Wing, S. (2013) Integrating informative priors from experimental research with Bayesian methods: An example from radiation epidemiology. *Epidemiology, 24*(1), 90-95. doi:10.1097/EDE.0b013e31827623ea


