

AMES, ALLISON JENNIFER, Ph.D. Bayesian Model Criticism: Prior Sensitivity of the Posterior Predictive Checks Method. (2015)  
Directed by Dr. Randall Penfield. 149 pp.

Use of noninformative priors with the Posterior Predictive Checks (PPC) method requires more attention. Previous research of the PPC has treated noninformative priors as always noninformative in relation to the likelihood, regardless of model-data fit. However, as model-data fit deteriorates, and the steepness of the likelihood's curvature diminishes, the prior can become more informative than initially intended.

The objective of this dissertation was to investigate whether specification of the prior distribution has an effect on the conclusions drawn from the PPC method. Findings indicated that the choice of discrepancy measure is an important factor in the overall success of the method, and that different discrepancy measures are affected more than others by prior specification.

BAYESIAN MODEL CRITICISM: PRIOR SENSITIVITY OF THE POSTERIOR  
PREDICTIVE CHECKS METHOD

by

Allison Jennifer Ames

A Dissertation Submitted to  
the Faculty of The Graduate School at  
The University of North Carolina at Greensboro  
in Partial Fulfillment  
of the Requirements for the Degree  
Doctor of Philosophy

Greensboro  
2015

Approved by

Dr. Randall Penfield  
Committee Chair

APPROVAL PAGE

This dissertation written by ALLISON JENNIFER AMES has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair

\_\_\_\_\_  
Dr. Randall Penfield

Committee Member

\_\_\_\_\_  
Dr. Robert Henson

Committee Member

\_\_\_\_\_  
Dr. Ric Luecht

Committee Member

\_\_\_\_\_  
Dr. John Willse

\_\_\_\_\_  
Date of Acceptance by Committee

\_\_\_\_\_  
Date of Final Oral Examination

## ACKNOWLEDGEMENTS

As a graduate student at UNCG I have received unwavering support and encouragement for this research from several individuals who I wish to acknowledge. Drs. Randall and Kara Penfield, who I put my faith in from beginning to end, are the best mentoring team, colleagues, and friends I could have hoped for. I would also like to thank my dissertation committee of Dr. Ric Luecht, Dr. Bob Henson, and Dr. John Willse for their support over the past several years in committee activities, the classroom, and as exemplars of the academic profession. In addition, other professors have provided invaluable advice and learning opportunities, including Dr. Micheline Chalhoub-Deville, Dr. Holly Downs, and Dr. Terry Ackerman. I would not be a graduate student at all were it not for Rachel Hill.

Other graduate students in the ERM department have played an integral role and it would be remiss of me not mention Nurliyana Bukhari, Robbie Furter, Jonathan Rollins, Kelli Samonte, and Emma Sunnassee for their contributions to my academic progress (and helping me laugh off the rest). Matt Boykin has also helped celebrate each success along the way and provided unconditional love during the process.

Finally, and most importantly, my family has provided me with the love I needed to finish. My parents, Dr. Glenn Ames and Kathryn Ames, have been my biggest cheerleaders. My brother, Pancho; sister in law, Amy; and nephews, Ethan and Cole, have provided a second home for me and a much-needed escape.

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	vii
LIST OF FIGURES .....	viii
CHAPTER	
I. INTRODUCTION .....	1
Definition of Key Terms Used in Bayesian Practice.....	9
Description of the Problem .....	11
Purpose .....	16
Research Questions .....	16
Organization of the Study .....	17
II. REVIEW OF THE LITERATURE.....	18
Bayesian Modeling Stages .....	19
The likelihood.....	19
Specification of priors.....	23
Construction of the posterior distribution.....	30
Inferences made from the posterior.....	31
Frequentist Approaches to Model-Data Fit .....	34
Bayesian Approaches to Model-Data Fit.....	40
Prior Specification in PPC .....	43
III. METHODS .....	52
Research Plan .....	53
Choice of Simulation Conditions .....	53
General.....	53
Introduction of Misfit.....	54
Data-generating parameters.....	59
Sample size.....	68
Discrepancy statistics.....	68
Choice of priors.....	69
Estimation .....	72
Outcomes .....	75
Analysis .....	75

IV. RESULTS.....	77
Convergence.....	78
Noninformative, Informative-Accurate, and Informative-Inaccurate Priors .....	80
Noninformative-Inaccurate Priors .....	80
Time to Converge .....	81
Analysis .....	83
Research Question 1: To What Extent Does Prior Specification Influence the Results of the PPC Method for the Model-data Fit of Unidimensional, Dichotomous IRT Models? .....	84
Hit Rate.....	84
<i>S-X<sup>2</sup></i> .....	85
<i>INFIT</i> .....	85
Percent correct. ....	85
False Positive Error .....	86
<i>S-X<sup>2</sup></i> .....	86
<i>INFIT</i> .....	87
Percent correct .....	87
Summary .....	88
Research Question 2: How Does Sample Size Affect the Influence of Prior Specification on the Results of the PPC Method for Model-data Fit of Unidimensional, Dichotomous IRT Models?.....	89
Hit Rate.....	90
<i>S-X<sup>2</sup></i> .....	92
<i>INFIT</i> .....	93
Percent correct. ....	94
False Positive Error .....	94
<i>S-X<sup>2</sup></i> .....	96
<i>INFIT</i> .....	97
Percent correct. ....	97
Summary .....	98
Research Question 3: How Does the Type of Misfit Affect the Influence of Prior Specification on the Results of the PPC Method for Model-data Fit of Unidimensional, Dichotomous IRT Models? .....	98
Hit Rates .....	99
<i>S-X<sup>2</sup></i> .....	101
<i>INFIT</i> .....	102
Percent correct. ....	103
Summary .....	103

Research Question 4: How Does the Interaction of Sample Size and Type of Misfit Affect the Influence of Prior Specification on the Results of the PPC Method for the Model-data Fit of Unidimensional, Dichotomous IRT Models?.....	104
Hit Rates .....	104
$S-X^2$ .....	104
$INFIT$ .....	104
Percent correct. ....	105
Summary .....	105
Other Considerations .....	105
Percent of misfitting items. ....	105
Summary .....	114
 V. DISCUSSION AND IMPLICATIONS .....	 118
Conclusions.....	118
Research question 1: Prior specification on PPC results. ....	118
Research question 2: Prior specification, sample size on PPC results.....	118
Research question 3: Prior specification, type of misfit on PPC results.....	119
Research question 4: Prior specification, sample size, type of misfit on PPC results.....	119
Implications.....	119
Item and Person Location .....	121
Type of Misfit .....	122
Sensitivity of $S-X^2$ to Prior Informativeness .....	123
Recommendations .....	124
Limitations .....	125
Directions for Future Research .....	126
 REFERENCES.....	 130

## LIST OF TABLES

	Page
Table 1. Applications of MCMC for IRT from 2009 – 2013 .....	3
Table 2. Generating Model Parameters in Sinharay, Johnson, and Stern (2006) .....	60
Table 3. MISFIT for GM=3PL and AM=2PL .....	61
Table 4. LPE-2PL Generating Model Parameters.....	64
Table 5. 3PL Generating Model Parameters.....	65
Table 6. 2PL Generating Model Parameters.....	67
Table 7. Prior Specification .....	71
Table 8. Simulation Conditions.....	72
Table 9. Mean Hit Rates by Prior Specification.....	84
Table 10. Mean False Positive Rates by Prior Specification .....	86
Table 11. Mean Hit Rates by Prior Specification and Sample size.....	90
Table 12. Effect Sizes for Research Questions Two Through Four.....	92
Table 13. Mean False Positive Rates by Prior Specification and Sample size .....	95
Table 14. Mean Hit Rates by Prior Specification, Sample size, and Type of Misfit .....	100
Table 15. Mean False Positive Rates by Type of Prior and Percent of Misfitting Items, for INFIT.....	106
Table 16. Mean Item Difficulty Values .....	122

## LIST OF FIGURES

	Page
Figure 1. Likelihood Curvature (dotted line) in Relation to the Prior (dashed line) and Effects on the Posterior (solid line) .....	15
Figure 2. IRFs for the Correct Option of Three Dichotomous Items .....	20
Figure 3. Illustration of Posterior Predictive Checks .....	41
Figure 4. Generating and Analysis Model Combinations.....	54
Figure 5. Types of Misfit in IRFs .....	55
Figure 6. IRFs for Hypothetical 2PL LPE Items ( $a_i = 1$ , $b_i = 0$ for all items) .....	58
Figure 7. Average Time to Converge, in minutes, by Type of Prior.....	82
Figure 8. Mean Hit Rates, by Type of Prior .....	88
Figure 9. Mean False Positive Rates, by Type of Prior .....	89
Figure 10. Mean Hit Rates, by Type of Prior, Sample size, and Discrepancy Statistic .....	91
Figure 11. Mean False Positive Rates, by Type of Prior, Sample Size, and Discrepancy Statistic .....	96
Figure 12. Mean False Positive Rates, by Type of Prior, Sample size, and Discrepancy Statistic .....	101
Figure 13. Mean False Positive Rates, by Type of Prior, Percent of Misfit, for INFIT .....	107
Figure 14. Hit Rates, with INFIT as Discrepancy Statistic, by Size of Misfit and Percent Misfit .....	109
Figure 15. Hit Rates, with INFIT as Discrepancy Statistic, by Location of Misfit and Percent Misfit .....	110
Figure 16. Hit Rates, with $S-X^2$ as Discrepancy Statistic, by Size of Misfit and Percent Misfit .....	111

Figure 17. Hit Rates, with $S-X^2$ as Discrepancy Statistic, by Location of Misfit and Percent Misfit .....	112
Figure 18. Hit Rates, with Percent correct as Discrepancy Statistic, by Size of Misfit and Percent Misfit .....	113
Figure 19. Hit Rates, with Percent correct as Discrepancy Statistic, by Location of Misfit and Percent Misfit .....	114

## CHAPTER I

### INTRODUCTION

Wainer (2010) has said that Bayesian methods are a suite of tools researchers must have in order to successfully tackle research problems looming in the future. 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). There has been a great emphasis on Bayesian methods of late (Ames & Samonte, in press; Andrews & Baguley, 2013). 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 and the Metropolis-Hastings Markov chain Monte Carlo (MCMC) algorithm should become second nature (Wainer, 2010).

Some have already heeded Wainer’s (2010) advice. The previous twenty years have seen a proliferation of studies using Bayesian methods in statistical research and publications, coinciding with advances in Monte Carlo methods. In fact, topics related to Bayesian methods now represent approximately 20% of published articles in statistics (Andrews & Baguley, 2013). This signifies a very important trend, or, more specifically, a paradigm shift. Andrews and Baguley, editors of a *British Journal of Mathematical and Statistical Psychology* special issue on the theory and practice of Bayesian statistics in psychology, comment that this shift will lead to an increased adoption of Bayesian

methods, which will have profound implications for the theory and application of data analysis. These implications will range in scope from journal editorial choices to how statistics is taught to university students (Andrews & Baguley, 2013).

This trend towards increasing numbers of Bayesian articles has also been witnessed in educational research, and, more specifically, item response theory (IRT) modeling. To illustrate this growing trend in the advancement of MCMC, a brief overview of the number and types of applications in IRT using MCMC methods in the past five years (2009 – 2013) is provided in Table 1.

The breadth of models found in Table 1 illustrates the broad use of Bayesian methods in educational methodology. MCMC methodology has been applied to dichotomous models (e.g., the two-parameter logistic model in Patz & Junker, 1999), polytomous models (e.g., the graded response model in Baldwin, Bernstein, & Wainer, 2009), and more complex models such as the cognitive diagnostic assessment fusion model (Jang, 2009), mixture Rasch with response time (Meyer, 2010), and logistic positive exponent (Bolfarine & Bazan, 2010), among others.

Table 1. Applications of MCMC for IRT from 2009 – 2013

<i>Author</i>	<i>Year</i>	<i>Journal</i>	<i>Model</i>	<i>I</i>	<i>N</i>	<i>Software</i>	<i>Burn-in</i>	<i>Iter</i>
Baldwin	2009	Statistics in Medicine	Graded response model	20	12	SCORIGHT	16500 0	20000
Bolfarine	2010	Journal of Educational and Behavioral Statistics	Logistic positive exponent	18	974	WinBUGS		
Choo	2013	Journal of Statistical Computation and Simulation	Mixture Rasch model	18	2156	WinBUGS		
Curi	2011	Statistical Methods in Medical Research	IRT for embarrassing items	20	348	WinBUGS		
De Gooijer	2011	Computational Statistics and Data Analysis	Two-parameter logistic (2PL)	6	25200			
de la Torre	2009	Applied Psychological Measurement	MIRT with ancillary information	77	1500	Ox	5000	20000
de la Torre	2009	Applied Psychological Measurement	Higher-order IRT	90	2255	Ox	3000	15000
de la Torre	2010	Applied Psychological Measurement	Higher-order IRT	90	2255	Ox	1000	10000

<i>Author</i>	<i>Year</i>	<i>Journal</i>	<i>Model</i>	<i>I</i>	<i>N</i>	<i>Software</i>	<i>Burn-in</i>	<i>Iter</i>
de la Torre	2009	Applied Psychological Measurement	IRT subscore methods	90	2255	Ox	2000	10000
de la Torre	2009	Journal of Educational and Behavioral Statistics	Deterministic-input noisy-AND (DINA) model	15	2144	Ox		
Edwards	2010	Psychometrika	Confirmatory item analysis	10 2	3000			
Entink	2011	Statistics in Medicine	Mixture multilevel IRT with survival model	30	668	BOA R package	5000	10000
Entink	2009	Psychometrika	Multivariate multilevel IRT	22, 65	286, 388	R	10000, 10000	5000, 20000
Finke	2009	Journal of Theoretical Politics	Two-parameter logistic (2PL)	61	82	GAUSS	10000	15000
Fragoso	2013	Biometrical Journal	Non-compensatory and compensatory MIRT two-parameter	21	1111		5000	100000
Fu	2009	Journal of Statistical Computation and Simulation	Multidimensional three-parameter logistic (3PL)	6	36	MATLAB	1000	15000
Fukuhara	2011	Applied Psychological Measurement	Bifactor MIRT for testlets	45	2000	WinBUGS	7000	15000

<i>Author</i>	<i>Year</i>	<i>Journal</i>	<i>Model</i>	<i>I</i>	<i>N</i>	<i>Software</i>	<i>Burn-in</i>	<i>Iter</i>
Geerlings	2011	Psychometrika	Hierarchical IRT for items in families	33 (11 families)	1350		20000	100000
Henson	2009	Psychometrika	Log-linear cognitive diagnosis	12	2144	MPLUS	5000	10000
Hsieh	2010	Multivariate Behavioral Research Applied	Generalized linear latent and mixed model	13	838	WinBUGS	4000	12000
Huang	2013	Psychological Measurement	Hierarchical IRT	24 7, 76	5000, 987	WinBUGS	1000	9000
Hung	2011	Multivariate Behavioral Research	Random-situation random-weight model with internal restrictions on item difficulty (MIRID)	10	268	WinBUGS	1000	4000
Hung	2010	Multivariate Behavioral Research	Multigroup multilevel categorical latent growth curve	7	264	WinBUGS	5000	10000
Hung	2012	Journal of Educational and Behavioral Statistics	Generalized multilevel facets model for longitudinal data	5	238	WinBUGS	8000	4000
Jang	2009	Language Testing Journal of	Fusion model (CDA)	37	2703	Arpeggio	13000	30000
Jiao	2013	Educational Measurement	One-parameter (1PL) testlet	54		WinBUGS	1000	2000

<i>Author</i>	<i>Year</i>	<i>Journal</i>	<i>Model</i>	<i>I</i>	<i>N</i>	<i>Software</i>	<i>Burn-in</i>	<i>Iter</i>
Jiao	2012	Journal of Educational Measurement	Multilevel testlet	32	1644	WinBUGS	2000	3000
Kang	2009	Applied Psychological Measurement	Polytomous models (focus on model selection indices)	5	3000	WinBUGS	5000	6000
Kieftenbeld	2012	Applied Psychological Measurement	Graded response model					
Kim	2009	Communications in Statistics in Medicine	SEM for ordinal response data w/ missingness	25	70548	FORTTRAN	1000	10000
Li	2012	Applied Psychological Measurement	Generalized Partial Credit Model	5	500		2000	152000
Li	2009	Applied Psychological Measurement	Mixture IRT	48	1200	WinBUGS	3000	10000
Luo	2013	Statistics in Medicine	Multilevel IRT	32	361	OpenBUGS	45000	50000
Meyer	2010	Applied Psychological Measurement	Mixture Rasch with response time	60	524	OpenBUGS	39999	30001
Saito	2010	International Journal of Methods in Psychiatric Research	Two-parameter logistic (2PL)	14	353	SAS/IML	1000	10000
Santos	2013	Journal of Applied Statistics	Skew multiple group IRT	20-80	295-568	Ox		

<i>Author</i>	<i>Year</i>	<i>Journal</i>	<i>Model</i>	<i>I</i>	<i>N</i>	<i>Software</i>	<i>Burn-in</i>	<i>Iter</i>
Soares	2009	Journal of Educational and Behavioral Statistics Applied	Integrated Bayesian for DIF	56	7998	MATLAB		
Stone	2009	Measurement in Education	Multidimensional IRT	59	10545	WinBUGS		
Tao	2013	Japanese Psychological Association	Two-parameter logistic testlet with testlet-level discrimination	28	1289		5000	30000
Usami	2011	Japanese Psychological Association	Generalized graded unfolding model	20	313	R	50000	50000
van den Hout	2010	Journal of the Royal Statistical society	Randomized response	3	2227	WinBUGS	50000	50000
Wang	2010	Statistics in Medicine	Testlet with covariates	21	718		10000	20000
Yao	2010	Journal of Educational Measurement	MIRT, bifactor	21 7	3953	BMIRT		

*Note.* The column titled *Author* represents first author only, *Year* is the year of publication, and *Journal* is the journal in which the study was published. Related to the MCMC modeling, the *Model* column of Table 1 is a brief description of the model estimated under the MCMC framework, *I* refers to number of items, *N* refers to number of respondents/subjects, *Software* is the software program implementing the MCMC algorithm, *Burn-in* is the number of burn-in iterations, and *Iter* is the number of MCMC samples drawn after the burn-in period.

Software for implementation of Bayesian estimation methods for IRT has also garnered increased attention. The development of multiple software packages, both commercial and open source, has aided the rapid adoption of Bayesian methods. Software to implement MCMC estimation has been described and illustrated in several sources, such as Curtis (2010), who provided BUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which contains code for common dichotomous and polytomous IRT models. Several ( $n = 17$ ) of the applications in Table 1 rely on BUGS software for conducting the MCMC analysis. Li and Baser (2012) provide a detailed explanation of how to perform MCMC estimation with the R (R Development Core Team, 2010) package R2WinBugs (Gelman, 2013) used in conjunction with BUGS. Working solely in the R environment, the package MCMCPack (Martin, Quinn, & Park, 2011) is another option for MCMC estimation. MPLUS (Muthén & Muthén, 2011) also has the capability to perform MCMC estimation, although not yet implemented in the applications found in Table 1. Yao (2003) developed BMIRT for MCMC of multidimensional IRT models. Recently, Ames and Samonte (in press) illustrated how to use SAS PROC MCMC (SAS Institute, Inc., 2014) for estimation of IRT models.

With several software options easily and freely available for MCMC implementation and the growing use of these software tools for IRT modeling, attention to the details of Bayesian methods and the potential associated pitfalls should be given considerable attention. These details include (a) specifying the likelihood model, (b) specifying the parameter for prior distribution(s), (c) obtaining the posterior distribution analytically using Bayes' Theorem or through sampling methods such as MCMC, and (d)

making appropriate inferences (O'Hagan, 1994). Each of these details can be considered a stage of the Bayesian modeling process and each has associated with it difficulties that may yet not have been sufficiently addressed. This research will address each stage of the modeling process, providing particular attention to both the second stage, prior specification, and a particular aspect of the inferential stage, that of model criticism.

Inevitably, the Bayesian versus Frequentist controversy enters into the discussion at some point with the investigation of Bayesian methods. One of the concerns raised by Frequentists is of interest to the remainder of this research, namely that Bayesian methods make subjective, rather than objective, claims. 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), which resulted in apprehensions 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. A brief introduction with more detail on the modeling stages is needed to describe the role of the prior and to facilitate the discussion of the potential subjectivity of the prior. What follows is a short discussion on these stages and general Bayesian terms.

### **Definition of Key Terms Used in Bayesian Practice**

All statistical probability models describe a mechanism, or relationship, that has generated the observed data as a function of unobserved parameters. One example is the interaction between people and items, resulting in dichotomous response data, often

modeled using IRT. The item parameters include difficulty or discrimination and the person parameters are defined by an individual's ability. However, information about the item parameters is not completely known, which introduces uncertainty into the relationship between the item response data and the item parameters. The fundamental difference between Bayesian and Frequentist methods lies in the treatment of this uncertainty. In the Bayesian paradigm, parameters are regarded as random variables (i.e., incorporate the uncertainty) and an entire distribution of possible parameter values is estimated, while Frequentist methods consider parameters as fixed, but unknown, quantities.

One of several modeling stages in the Bayesian approach is the specification of a model for the observed data. This model simply describes the process giving rise to the data in terms of the unknown parameters (O'Hagan, 1994). The model for observed data is often termed the *likelihood*. The likelihoods observed in recent IRT applications can be found in Table 1 in the *Model* column.

Specification of parameter uncertainty before any data are observed is another stage in Bayesian modeling. This uncertainty specification is termed the *prior* information because it represents information known about the parameter before observing any data. Prior beliefs could be *elicited*, coming from other observed data and research or representing expert opinions from the field (Fox, 2010). Alternatively, the prior information could reflect a relative lack of understanding surrounding the parameters. This type of information is considered *noninformative*. These noninformative priors are those in which nearly, or completely, identical probability is given to all

possible parameter values (Lambert et al., 2005). Noninformative priors are often termed *flat* and used so that the data and likelihood drive the estimation procedure rather than the prior.

In the next stage, after the data are observed, prior information on the parameters is combined with the likelihood to provide a distribution of parameter information. This combination of the likelihood and priors comes via Bayes' Theorem. Because the combination occurs after the data are observed, the distribution of potential parameter values is known as the *posterior* parameter distribution (Fox, 2010). Posterior distributions express what is known about parameters once the data has been observed. They specify the probability that each parameter equals a particular value, or lies in a certain range of values.

In the final stage, the inferential stage, both the likelihood and priors must be assessed in terms of their appropriateness to the data at hand (Gelman & Shalizi, 2013). Misspecification of either or both of the likelihood and the priors could lead to inappropriate inferences (Evans & Moshonov, 2006). Several approaches for Bayesian evaluation of model-data fit exist, including (a) examining the sensitivity of inferences to changes in the prior distribution and the analysis model, (b) checking the sensibility of posterior inferences against the researcher's substantive knowledge, and (c) checking that the model fits the data (Gelman, Meng, & Stern, 2003).

### **Description of the Problem**

The majority of the focus of Bayesian IRT model-data fit has come via the third approach, introduced above, termed *posterior predictive checks* (PPC; Gelman et al.,

2003), which consists of assessing the plausibility of data simulated from the posterior against the observed data. However, while PPC has proven to be a useful tool for evaluating fit of IRT models (Sinharay, 2006), specifically the likelihood component, there has been little consistency in the implementation of simulation studies in the area when applied to IRT. For instance, several studies (Sinharay, 2005; Sinharay & Johnson, 2003; Sinharay, Johnson, & Stern, 2006) have used noninformative priors for dichotomous models. Sinharay (2006) used informative priors with similar simulation conditions to his other studies in the area. Ames (2014) used noninformative priors for PPC applied to polytomous models. In contrast, Zhu and Stone (2011) used more narrowly focused priors for the graded response model (GRM; Samejima, 1969), which were based on asymptotic parameter distributional assumptions.

Bayesian parameter recovery studies have also varied the use of priors. For instance, Kieftenbeld and Natesan (2012) used noninformative priors for parameter recovery of the GRM with good model-fit data. Another parameter recovery study (Wollack, Bolt, Cohen, & Lee, 2002) used a different version of noninformative priors for recovery of item parameters in the nominal response model (again with good model-fit data).

Use of noninformative priors also has come under criticism for several other reasons. It has been recommended that noninformative priors should be used only as a placeholder in Bayesian methods rather than a final prior specification, that is, the researcher can use the noninformative prior to get the analysis started but if the resulting posterior inferences lack precision, or do not make sense, the prior information should be

modified (Gelman & Shalizi, 2013). Additionally, previous research (Gelman et al., 1996; Zhu, 2009) has also shown situations in which the use of noninformative priors has been too strong, with the result being that the noninformative priors represented very strong prior information instead of the intended role of having minimal influence.

An additional criticism of noninformative priors stems from the previously discussed parameter recovery studies (Kieftenbeld & Natesan, 2012; Wollack et al., 2002). Kieftenbeld and Natesan (2012) found that almost all item discrimination parameters were positively biased, likely due to the mean of the noninformative prior being larger than any of the generating discrimination values. In their study, good model-data fit was assumed, as the data-generating model was the same as the data-analysis model. However, despite the good model-data fit, Kieftenbeld and Natesan (2012) found that the noninformative priors likely induced positive bias in some item parameter estimates, which could be potentially problematic when applying the PPC method. Because the data are simulated from misspecified posteriors, then the simulated data will likely be systematically dissimilar to the observed data and the PPC method will find evidence of misfit due to the misspecified posterior. Because the data were generated according to the same likelihood model to which the data were fit, the finding of model-data misfit would be incorrect.

This brings to light a critical point surrounding the definition of noninformative and use of this type of distribution for a prior. A prior distribution should be considered noninformative only if the prior is flat relative to the likelihood function. If the likelihood

is flat relative to the prior, then the prior will still be relatively informative despite the researcher's attempts to let the likelihood and the data drive the posterior estimation.

In IRT, aberrant response patterns across items or across people will tend to result in a flatter likelihood (Drasgow, Levine, & McLaughlin, 1991). These aberrant response patterns represent a form of model-data misfit. The flatter the likelihood, the less likely that a flat prior will have a minimum role on posterior inferences as is intended (O'Hagan, 1994). Therefore, the use of noninformative priors in PPC studies may have, in fact, introduced strong prior information, which could influence the model-data fit conclusions regarding the likelihood.

Figure 1 illustrates the flatness of the likelihood and the role of the prior as the likelihood becomes less curved. In the four panels, in each case the prior is held constant (dashed line). The curvature of the likelihood is varied (dotted line) and the resulting posterior (solid line) changes accordingly. In the top left panel, the likelihood is almost flat and the prior, while specified noninformative, drives the posterior construction as much as the likelihood does. As the likelihood becomes less flat in relation to the prior, the likelihood begins to shape the posterior's shape and curvature. In the bottom right panel, the posterior's construction is driven almost entirely by the likelihood, which is very peaked in relation to the prior.

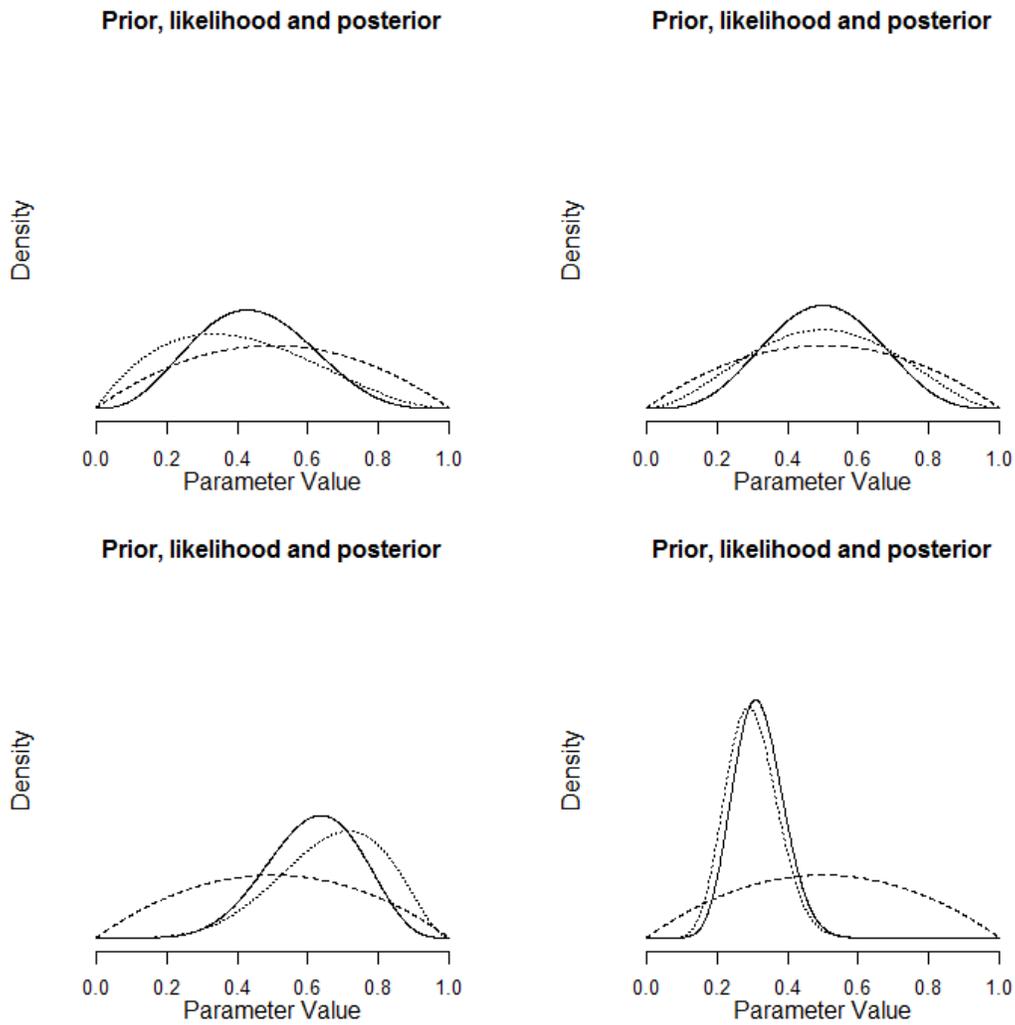


Figure 1. Likelihood Curvature (dotted line) in Relation to the Prior (dashed line) and Effects on the Posterior (solid line)

Thus, despite specification of a noninformative prior, the flatness of the likelihood relative to the prior is the determining factor in the shape of the posterior. It does not appear that this has been considered in PPC for IRT studies. The PPC studies applied to IRT have only discussed flatness of the prior and have not provided a discussion on the curvature of the likelihood in relation to the curvature of the prior. Further, whenever

surprising or aberrant data are observed, sensitivity of posterior inferences to prior specification should be suspected (O'Hagan, 1994). Thus, specification of the prior in PPC studies requires careful consideration and it might be the case that using more informative priors would be useful for PPC studies.

### **Purpose**

Use of noninformative priors with the PPC method requires more attention. Previous research of the PPC has treated noninformative priors as always noninformative in relation to the likelihood, regardless of model-data fit. However, as model-data fit deteriorates, and the steepness of the likelihood's curvature diminishes, the prior can become more informative than initially intended.

Further research is required to determine which priors best reflect inconsistencies between data and the null hypotheses and will best detect model-data misfit using the PPC method (Berkhof, van Mechelen, & Hoijtink, 2000). The objective of this study is to investigate whether specification of the prior distribution has an effect on the conclusions drawn from the PPC method regarding model-data fit. Specifically, the following four research questions will be addressed.

### **Research Questions**

1. To what extent does prior specification influence the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?
2. How does sample size affect the influence of prior specification on the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?

3. How does the type of misfit affect the influence of prior specification on the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?
4. How does the interaction of sample size and type of misfit affect the influence of prior specification on the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?

### **Organization of the Study**

To answer the four research questions, a review of the relevant literature on PPC and prior sensitivity will be provided as well as a detailed proposal of the methodology to be used in the study. Chapter Two reviews the literature in the area, beginning with a discussion of Bayesian and PPC methods, specifically of work related to Bayesian IRT modeling and prior specification in these studies. Chapter Three outlines appropriate methodologies to answer each of the four research questions, including operationalization of the variables and details of the analysis methods which will be used to answer the research questions.

## **CHAPTER II**

### **REVIEW OF THE LITERATURE**

Bayesian estimation methods require specification of a likelihood model as well as specification of parameter priors (Gelman, Carlin, Stern, & Rubin, 2003). 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.

Misspecification of either the likelihood or the priors could lead to inappropriate inferences (Evans & Moshonov, 2006; O'Hagan, 1994). The extent to which inferences are sensitive to prior or likelihood misspecification is an important consideration in a Bayesian analysis (O'Hagan, 1994). In IRT, specification of the likelihood has been checked by PPC. However, little attention has been paid to the potential sensitivity of the PPC method to a misspecified prior parameter or the appropriateness of the prior itself.

What follows in this chapter is a discussion of the details of the Bayesian modeling stages, which will be used to facilitate discussion of the literature regarding prior specification and the PPC method. This chapter begins with a discussion of Bayesian modeling stages, including likelihood specification, prior specification, construction of the posterior via MCMC, and making inferences. The final stage of making inferences involves the PPC method. Following the Bayesian model stages, examination of the potential confounding of prior specification and the PPC method will be described.

## Bayesian Modeling Stages

**The likelihood.** Assume the observed item response data,  $x$ , are used to measure some parameter,  $\theta$ . To express the prior information on  $\theta$ ,  $f(\theta)$  is used. The likelihood function of  $x$  is denoted by  $f(x|\theta)$  and the posterior of  $\theta$  is denoted by  $f(\theta|x)$ , representing the conditional distribution of both: (a) parameters given the prior beliefs and (b) the observed data. Bayes' Theorem provides the link between the prior, likelihood, and posterior and is represented via

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta}. \quad (1)$$

The likelihood function,  $f(x|\theta)$ , describes the probability of observing the item response data given the parameter values. In the Bayesian framework, the data are assumed to affect the posterior inference only through the likelihood (Gelman et al., 2003).

This study will focus on the IRT likelihood models, with an emphasis on models for unidimensional dichotomous items. In this proposal,  $i$  will represent the item of interest ( $i = 1, 2, \dots, I$  items) and  $n$  will represent the examinee ( $n = 1, 2, \dots, N$  examinees). The examinee's latent ability trait will be denoted by  $\theta$ . Item responses will be denoted by  $x$ , with the  $n^{\text{th}}$  examinee's response to the  $i^{\text{th}}$  item denoted by  $x_{ni}$ . Thus,

$$x_{ni} = \begin{cases} 1, & \text{if examinee } n \text{ answers item } i \text{ correctly} \\ 0, & \text{if examinee } n \text{ answers item } i \text{ incorrectly} \end{cases}$$

Two fundamental concepts of IRT are that the performance of an examinee on an item can be predicted by  $\theta$  and that there is a link between an examinee's response to an item and  $\theta$  (Hambleton, Swaminathan, & Rogers, 1991). The latter concept, the link between item responses and  $\theta$ , is referred to as the *item response function* (IRF). The IRF for  $x_{ni} = 1$  specifies the probability of correct response as a function of  $\theta$ . The shape of the IRF implies that individuals with higher levels of the trait should have a higher chance of getting the item correct than individuals with lower levels of the trait. An example of this can be found in Figure 2. Looking at Item 1 in Figure 2 (solid black line), as the individual's  $\theta$  increases, so too does the height of the curve, which represents the probability of correct response.

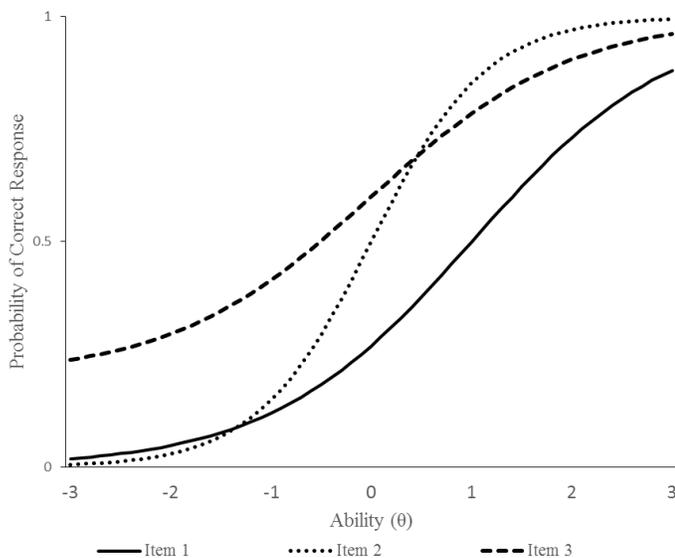


Figure 2. IRFs for the Correct Option of Three Dichotomous Items. For Item 1,  $a_1 = 1$ ,  $b_1 = 1$ ,  $c_1 = 0$ ; for Item 2,  $a_2 = 1.75$ ,  $b_2 = 0$ ,  $c_2 = 0$ ; and for Item 3,  $a_3 = 1$ ,  $b_3 = 0$ ,  $c_3 = 0.2$ .

The particular shape and location of an IRF reflects the psychometric properties of the item, such as difficulty, discrimination, and guessing (Hambleton et al., 1991; Lord, 1980). The different forms of the three IRFs shown in Figure 2 reflect differences in the items' difficulty, discrimination, and guessing. Location of the curve is dictated by the item's difficulty parameter. The IRF for Item 1 reflects the highest level of difficulty. Steepness of the curve is dictated by the item's discrimination parameter. The IRF for Item 2, with the steepest slope, reflects the highest degree of discrimination. The lower asymptote of the curve – the lowest probability of correct response an examinee has for answering the item correctly – is dictated by the guessing parameter. The IRF for Item 3, which has a lower asymptote near 0.2, reflects the highest degree of guessing.

The IRF for  $x_{ni} = 1$  is defined by the mathematical model. In MCMC studies, there are two model types commonly used for IRT: the normal ogive models (often used due to their desirable mathematical properties), and special cases of the general logistic regression model. The two forms of IRT models closely resemble each other when the logistic item parameter values are multiplied by a constant scaling factor of 1.7. With this scaling, for different levels of  $\theta$ , the response probabilities differ by no more than 0.01 (Hambleton et al., 1991).

A simple IRF for  $x_{ni} = 1$  is the one-parameter normal ogive (1PNO; Lord and Novick, 1968) model in which the IRF is based on a cumulative normal distribution. The 1PNO models probability of correct response via

$$P(x_{ni} = 1 | \theta_n, \delta_i) = \Phi(\theta_n - \delta_i), \quad (2)$$

where  $\delta_i$  is the item's difficulty measure,  $\theta_n$  is an examinee's scalar latent ability, and  $\Phi(\cdot)$  is the cumulative normal distribution function. Item discrimination is introduced through  $\alpha_i$ , a positive scalar, and the two-parameter normal ogive (2PNO) is denoted by

$$P(x_{ni} = 1 | \theta_n, \alpha_i, \delta_i) = \Phi(\alpha_i \theta_n - \delta_i). \quad (3)$$

The three-parameter normal ogive (3PNO) is denoted by

$$P(x_{ni} = 1 | \theta_n, \alpha_i, \delta_i, \varphi_i) = \varphi_i + (1 - \varphi_i) \Phi(\alpha_i \theta_n - \delta_i), \quad (4)$$

where  $\varphi_i$  is the lower asymptote for the IRF, bounded between 0 and 1.

Similar to the 1PNO is the one-parameter logistic (1PL), where the IRF for  $x_{ni} = 1$  is modeled by

$$P_{ni}(x_{ni} = 1 | \theta_n, a, b_i) = \frac{\exp(a(\theta_n - b_i))}{1 + \exp(a(\theta_n - b_i))}. \quad (5)$$

where  $b_i$  is the item's difficulty measure, and item discrimination is denoted by  $a$ .

The two-parameter logistic (2PL) is represented by

$$P_{ni}(x_{ni} = 1 | \theta_n, a_i, b_i) = \frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))}. \quad (6)$$

The three-parameter logistic (3PL; Birnbaum, 1968) is akin to the 3PNO and is a more general model for dichotomous responses. The 3PL is represented by the statistical IRT model

$$P_{ni}(x_{ni} = 1|\theta_n, a_i, b_i, c_i) = c_i + (1 - c_i) \frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))}, \quad (7)$$

where  $c_i$  is the lower asymptote for the IRF, bounded between 0 and 1. For simplicity in notation, the IRF for  $x_{ni} = 1$  will be shortened to  $P_{ni1}$  for the remainder of this text. For example, for the 3PL in Equation ,  $P_{ni}(x_{ni} = 1|\theta_n, a_i, b_i, c_i) = P_{ni1}$ .

The models described in Equations 2 - 7 provide a probabilistic model of the data, each with a set of underlying assumptions. Each IRT model is *falsifiable*, indicating the model may or may not be appropriate for the data (Hambleton et al., 1991). Assessment of these assumptions is crucial to ensuring that valid inferences are drawn from the IRT models (Hambleton et al., 1991; Kang & Chen, 2008). When an IRT model demonstrates poor adherence to the assumptions, several undesirable outcomes are possible, such as biased ability and item parameter estimates (Wainer & Thissen, 1987; Yen, 1981). These consequences of the model not adhering to its assumptions complicates the application of IRT models in such areas as test development, equating, and computer adaptive testing (Kang & Chen, 2008). More detail on the methods for assessing appropriateness of model assumptions is found in the section on drawing inferences, but it is crucial that the likelihood model of interest be subject to rigorous testing to ensure the relevant model assumptions are upheld and valid inferences are being made.

**Specification of priors.** 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, whereas the Frequentist paradigm assumes a single, fixed point value for the parameter.

The prior distribution can be viewed as an assumption of the model (Gelman & Shalizi, 2013), similar to the assumptions underlying the likelihood. Similar to the likelihood assumptions, prior assumptions must also be assessed for their appropriateness. Specification and assessment of prior distributions may result in the decision that they are incorrectly specified. The prior specification may then need to be revised, revisited, or completely thrown out. As with evaluating the IRT model, criticism of priors is based upon their suitability to the data being studied (Gelman & Shalizi, 2013).

Recognizing that the prior distribution is a testable part of the whole Bayesian model is a critical part of prior specification (Gelman & Shalizi, 2013). One sphere of thought is that there are unique, objectively correct prior distributions for each situation (Jaynes, 1968). However, attempting to devise these has proven unproductive (Kass & Wasserman, 1996). Given this lack of correct distributions, it is more likely that practitioners will choose among some general classes of prior distributions

The researcher has several options for incorporating the prior information into the Bayesian modeling process (Albert & Louis, 2000). *Elicited priors* are those in which the information is elicited from experts with information about the substantive question of interest, but who are not involved in the model construction process. These elicited priors could also arise from a collection of possible values of the parameter informed sequentially through previous studies in the area. For instance, suppose two independent

samples are collected from year 1 and year 2. The posterior for the full data can be obtained by first finding the posterior for the first year's data set and then using this posterior as the prior for the second year's data (Albert & Louis, 2000).

Prior specification could also arise from common distributional families, such as the Normal or Gamma distributions, which are tied to distributional assumptions. These are also considered elicited priors. As an example, latent ability estimates from unidimensional IRT models are often considered to come from a Normal( 0 , 1 ) distribution to avoid indeterminacy of the parameterization of several IRT models (Lord, 1980; Tsutakawa, 1992). Elicited priors could be strong and narrowly focused, or they could be weak, reflecting a less focused range of inference, but still with some informative qualities. Typically, the strength of a prior distribution is controlled by the distribution's variance, with smaller variances demonstrating more strength and prior *precision*. For instance, a Normal( 0 , 1 ) prior distribution would be considered more precise and stronger than a Normal( 0 , 100 ) prior distribution.

In the case where the posterior distributions are in the same distributional family as the prior distribution, the prior is called a *conjugate prior* (Raiffa & Schlaifer, 1961). The benefit of these conjugate priors is that they can help permit posterior distributions to emerge without numerical integration, a decided benefit for practitioners looking to avoid complicated integrals. Conjugate priors play an important role in estimation of IRT models through MCMC methods. They are desirable in that the use of a conjugate prior results in a posterior distribution of a known functional form, and, thus make sampling in MCMC more computationally efficient (Kim & Bolt, 2007). An additional benefit is that

the posterior predictive distribution (PPD; the distribution of future model data) of an exponential-family random variable with a conjugate prior can be written in closed form (Gelman et al., 2003).

Another option for the specification of the prior is the use of *noninformative priors*. In the Bayesian paradigm, noninformative priors are recommended when no reliable information about the parameter exists (Albert & Louis, 2000) or if maximum likelihood estimates of item or person parameters are desired (Patz & Junker, 1999). However, use of noninformative priors negates many Bayesian method advantages by essentially reducing the estimation solution to a maximum likelihood estimate (Fox, 2010). Further, use of noninformative priors implies that the posterior arose from the data only and that all resulting inferences were completely objective rather than subjective.

A closely related notion to the noninformative prior is that of the *reference prior*. These are treated as a convenient place to begin an analysis (Kass & Wasserman, 1996; Carlin & Louis, 2000). An example of a reference prior for IRT modeling could be the approach wherein all items are assigned a difficulty value of a standard Normal distribution and then the prior's performance is evaluated under a prior sensitivity analysis framework. If changes to the prior are deemed warranted, the reference prior can be abandoned or modified.

Prior distributions specified as *uniform* are often used as both reference and noninformative priors. These uniform priors are often termed *flat*, as they indicate the value of the parameter is equally likely across the specified range. However, Carlin and Louis (2000) show that the uniform prior is not invariant under reparameterization.

Another problem with the use of a flat prior is that they are usually *improper*, meaning they cannot be normalized to integrate to a value of one, a requirement for all probability densities. The improper prior could lead to an improper posterior, resulting in invalid inferences (Ghosh, Ghosh, Chen, & Agresti, 1997).

An alternative to the flat prior is the *Jeffreys prior* (Jeffreys, 1961), which is invariant under reparameterization and results in a proper posterior. The Jeffreys prior is a reference prior and defined for  $\theta$  as the prior which maximizes the empirical usefulness of a test (Markon, 2013). As the number of items on a test increases towards infinity, test information does not increase evenly across the range of  $\theta$ , but rather increases proportional to the Fisher information function. The shape of the Fisher information function is used to define the Jeffreys prior for  $\theta$ , allowing the placement of maximum prior probability on regions of  $\theta$  parameters where the test has optimal power to distinguish among examinees (Markon, 2013).

When using a Jeffreys prior, the form of the likelihood helps to determine the prior because the Jeffreys prior is proportional to the square root of the test information function, specified via

$$P^*(\theta) = \frac{\sqrt{I(\theta)}}{\int \sqrt{I(\theta)}}, \quad (8)$$

where  $I(\theta)$  is the equation for the Fisher's information function. Information is proportional to the second derivative of the log-likelihood function, which reflects the curvature of the log-likelihood at a particular value of  $\theta$ .

Multiple researchers have provided suggestions on choosing priors for IRT models. The suggestions have been quite varied. For instance, when Bayesian estimation and MCMC was first applied via Gibbs sampling to IRT, Albert (1992) used noninformative priors. Kim and Bolt (2007) suggest noninformative priors as well, indicating their use is appropriate if the marginal maximum likelihood solution is desirable. However, the specification provided for discrimination parameters was chosen to reflect that of the IRT calibration software PARSCALE (Muraki & Bock, 1997), and the priors in Kim and Bolt (2007) are more informative than many other so-called noninformative specifications (e.g.,  $a_i \sim \text{lognormal}(0, 0.5)$ ). Conversely, Begun and Glas (2001) and Glas and Meijer (2003) used very precise priors for the 3PNO. Tsutakawa (1992) reparameterized the 3PL, specifying the probability of correct response for any three ordered points on the ability scale. He then proposed using a constrained Dirichlet prior distribution on these probabilities. This is in direct contrast to Patz and Junker (1999) who assumed that the parameters are a priori independent of one another, using standard normal priors for examinee ability and item difficulty, and lognormal priors for discrimination parameters.

Sheng (2010) investigated the role of prior specification on parameter recovery of 3PNO IRT models. The informativeness of the priors was manipulated by the variance of the prior distribution, with the mean of each prior distribution held constant at 0. Informativeness of the prior affected not only convergence rates, the rate at which the MCMC algorithm closed in on the posterior distribution of the parameter given the data, but the precision and accuracy of parameter posterior point estimates as well. This was

particularly true for small sample sizes and small numbers of items. Sheng also noted that small prior variances result in Bayesian estimates that are closer, or shrunken in, towards the prior mean than those estimates resulting from larger prior variances. Thus, if the prior distribution is appropriately informative, the Bayesian point estimates will be less likely to take on unreasonable values. Sheng's findings indicate that for the 3PNO, appropriately specified informative priors should be adopted for discrimination and difficulty parameters to obtain more efficient and accurate parameter estimates with small samples and/or short tests. With little prior information, informative priors were not recommended.

Howell and Janosky (1991) found a similar feature with the 2PL. With large data sets, the informativeness of the prior distribution had little effect on parameter estimates. However, with small samples and/or short exams, informative and appropriately specified priors can have a profound effect on the resulting MCMC estimates. The benefits of informativeness are contingent on appropriately specified prior distributions. With informative but inaccurate priors, biased estimates and incorrect posterior inferences can result (Mislevy, 1986).

The above parameter recovery studies were undertaken with the expectation of good model-fit data, but just as much variation in prior specification has been witnessed in the presence of poor model-fit data. When trying to detect misfit via the PPC method, several studies (Sinharay, 2006; Sinharay, Johnson, & Stern, 2006; Sinharay, 2005; Sinharay & Johnson, 2003) have used noninformative priors for dichotomous models ( $\log(a_i) \sim \text{Normal}(0, 10)$  and  $b_i \sim \text{Normal}(0, 10)$ ) in the presence of simulated misfit.

In contrast, Sinharay (2006) used more informative priors ( $\log(a_i) \sim \text{Normal}(0, 1)$  and  $b_i \sim \text{Normal}(0, 1)$ ). Ames (2014) used noninformative priors for PPC applied to polytomous models. In contrast, Zhu and Stone (2011) used more narrowly focused priors for the GRM (e.g.,  $\log(a_i) \sim \text{Normal}(0, 1)$  and  $b_{ik} \sim \text{Normal}(0, 1)$ ). Another study on dichotomous items (Toribio & Albert, 2011) failed to provide any information on the prior parameters used. However, regardless of the prior specification, the prior must be evaluated as other model assumptions are; several researchers have even stressed that, with Bayesian analysis, it is important for interpretation of results to disentangle the role of the likelihood and the prior on the posterior distribution (Muller, 1989).

**Construction of the posterior distribution.** Often, estimating the posterior is not directly possible via Bayes' Theorem because of the analytical complexity involved in integrating the complex IRT likelihoods. MCMC, which is 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. The Monte Carlo principle states that anything a practitioner wishes to know about a random variable can be learned by sampling many times from the probability distribution of that random variable (Jackman, 2009). Therefore, if the practitioner wishes to learn about the posterior of  $\theta$ , they must sample many times from the posterior of  $\theta$ . If the posterior is a common distribution (e.g., Normal), sampling from the posterior is relatively simple because it has a known form. However, the posterior is often not a simple distribution and MCMC is then needed in order to learn about the posterior.

A Markov chain (a sequence of random variables) is used to sample from an unknown distributional form. The sequence has a particular property: the random variable at the current time depends only on its immediate predecessor (i.e., if  $\theta$  represents the random variable, then the value of  $\theta$  in time  $t$  of the sequence depends only on the value of  $\theta$  in time  $t-1$ ). Once constructed, the Markov chain represents the posterior distribution and each point in the chain represents a sample of the posterior.

The posterior can be thought of as a combination of the prior and the likelihood. Intuitively, the posterior is determined by the amount of information contained in both the likelihood and the prior. In general, if the prior information is weak relative to the information contained in the likelihood, then the posterior will be relatively unaffected by the form of the prior and by prior misspecification (O'Hagan, 1994). Inferences drawn from a posterior would be relatively unaffected by the prior misspecification in this case. Similarly, if likelihood information is weak relative to the information contained in the prior, then the posterior will be relatively unaffected by the form of the likelihood and less vulnerable to likelihood misspecification. Inferences drawn from a posterior would be relatively unaffected by this type of misspecification (O'Hagan, 1994).

**Inferences made from the posterior.** This section concerns the inferences drawn from the parameter posterior distribution constructed via MCMC methods. Inferences can be made once the posterior is constructed. Each of the examinee ability and item parameter posterior distributions provides complete information about the associated parameter (Fox, 2010). In Bayesian inference, the posterior is often summarized by several statistics and figures, providing information on where most of the posterior mass

is located (Fox, 2010) and these summaries often include the mode, mean, and standard deviation of the posterior distribution.

The extent to which inferences made from parameter posteriors are sensitive to prior and/or likelihood misspecification is an important consideration in a Bayesian analysis (O'Hagan, 1994). The likelihood component, the IRT model of interest, must be checked for misspecification - that is, their adherence to model assumptions needs to be deemed adequate before inferences can be validly drawn. Misspecification testing can take place in a Frequentist framework (see Spanos, 2007, for error testing and misspecification; Ames and Penfield, under review, for a review of commonly used Frequentist approaches) as well as a Bayesian framework.

The IRT model assumptions to be checked in misspecification tests are (a) unidimensionality, (b) local independence, and (c) that the probabilistic model reflects the true link between latent traits and item responses (Hambleton et al., 1991). The first assumption, *unidimensionality*, specifies that only one latent trait is measured by the items on the test. Typically, this assumption is not met in an absolute sense because other factors interact with an examinee's response to an item such as test anxiety and other cognitive skills (Hambleton et al., 1991). What is required for the unidimensionality assumption to be adequately met is the presence of a single, dominant trait being measured on the test. For instance, an item purporting to measure algebra ability should be dominated by the latent algebra ability trait, and not by an examinee's reading ability. Several multidimensional IRT models have been proposed, allowing for the presence of

more than one latent trait. For a review of multidimensional models, see van der Linden and Hamilton (Section III, 1997).

Another assumption of IRT models is that of *local independence*, meaning that, after conditioning on examinee ability, responses to any particular item are statistically independent (Hambleton et al., 1991). When the assumption of unidimensionality holds true, local independence is automatically obtained (Lord, 1980; Lord & Novick, 1968). The property of local independence yields the result that, for a given response pattern, the probability of that pattern on a set of items is equal to the product of the probability of responses to the individual items (Hambleton et al., 1991). This results in dichotomous models having the following joint likelihood

$$Likelihood = L = \prod_{n=1}^N \prod_{i=1}^I P_{ni1}^{x_{ni}} (1 - P_{ni1})^{(1-x_{ni})}, \quad (9)$$

and log-likelihood

$$Log - likelihood = LL = \sum_{n=1}^N \sum_{i=1}^I [x_{ni} * \log(P_{ni1}) + (1 - x_{ni}) * \log(1 - P_{ni1})], \quad (10)$$

where  $P_{ni1}$  represents the IRF for the particular model under consideration.

The final assumption is that the IRF reflects the true link between latent traits and item responses. This is often referred to as model-data fit. When the IRF reflects the true link between latent traits and item responses, the IRF represents the same link, regardless of the group of examinees responding to the item. That is, the IRF is *invariant* across the groups of examinees. When the IRF does not reflect the true link between latent traits and

item responses, the model will not yield invariant item and ability parameters (Hambleton et al., 1991).

Beyond a lack of invariance, another consequence of model-data misfit is seen in the informativeness of the likelihood. The weaker the amount of information in the data, the flatter the distributional shape and the less informative the likelihood will be (Drasgow, Levine, & McLaughlin, 1991). Weak information could come from aberrant response patterns across people or across items. These aberrant response patterns would flag an item as “misfitting” using traditional model-data misfit statistics. Another term also used to describe data which are a good fit to the given model is *strong data*. These strong data result in a likelihood that is quite “peaked,” with negligible information outside of a small region around its maximum (O’Hagan, 1994). Weak data present the opposite: the likelihood is less-peaked and there is less information at the maximum.

Assessment of how well the true link is represented by the model is, therefore, an important task. One method of assessing model-data fit, and, therefore, the informativeness of the likelihood, is through an examination of predicted responses compared to actual, observed responses. Frequentist methods provide an important element in the Bayesian model criticism process. Further, the Bayesian procedures can be conceptualized as an extension of the Frequentist approach.

### **Frequentist Approaches to Model-Data Fit**

Suppose the IRF posits a high probability of correct response for an examinee resulting in the prediction that the examinee will answer the item correctly. If the examinee does answer correctly, the prediction is accurate. However, if the examinee

does not answer correctly, the prediction would be considered inaccurate. Frequentist methods for evaluating IRT model-data fit are based on examining how closely the observed responses fit those predicted by the model. As the difference between observed and predicted data increases, the fit of the model to the observed data decreases, which provides evidence of misfit.

The difference between the observed response and the predicted response is termed the *residual* ( $r_{ni}$ ), and is denoted by

$$r_{ni} = x_{ni} - P_{ni1}. \quad (11)$$

One limitation of interpreting fit using the individual-level residuals shown in Equation 11 is that the individual-level residuals will typically not be zero, even in the presence of relatively good model-data fit. As a result, individuals are grouped according to specific ranges of ability, referred to as *bins*, and then the difference between the observed proportion correct for each bin and the proportion correct predicted for each bin is considered. The bin-level residual for bin  $h$  is given by

$$r_{hi} = O_{hi1} - P_{hi1}, \quad (12)$$

where  $O_{hi1}$  represents the observed proportion of individuals in the  $h^{\text{th}}$  bin having a correct response to the  $i^{\text{th}}$  item and  $P_{hi1}$  is the probability of correct response for that bin. The bins will be denoted here by  $h = 1, 2, \dots, H$ , such that  $H$  represents the total number of bins.

Several different statistical approaches have been developed for evaluating fit, all of which involve the concept of the residual. Despite the voluminous options, most follow one of two general approaches: a chi-square approach and a likelihood-ratio approach. The chi-square approach to evaluating model-data fit in dichotomous items is given by the general form of

$$\chi_i^2 = \sum_{h=1}^H N_{hi} \frac{(r_{hi})^2}{P_{hi1}(1-P_{hi1})}, \quad (13)$$

where  $N_{hi}$  represents the number of people in bin  $h$  responding to item  $i$ . The chi-square statistic in Equation 13 represents the sum of squared, standardized residuals. The chi-square approach embeds the bin-level residuals directly so that as the residuals increase, so too do the values of the chi-square statistic. Examples of this statistic are Yen's  $Q_1$  (1981) and Bock's  $\chi^2$  (1960). The likelihood-ratio based statistic for dichotomous items is given by the form

$$LR_i = 2 \sum_{h=1}^H \left[ N_{hi1} \ln \left( \frac{N_{hi1}}{N_{hi} P_{hi1}} \right) + N_{hi0} \ln \left( \frac{N_{hi0}}{N_{hi} (1-P_{hi1})} \right) \right], \quad (14)$$

where  $N_{hi1}$  and  $N_{hi0}$  represent the number of people in bin  $h$  answering item  $i$  correctly and incorrectly, respectively. Some minor algebraic manipulation reveals that the natural log of bin-level residuals is involved. The term  $\ln \left( \frac{N_{hi1}}{N_{hi} P_{hi1}} \right)$  expands to  $\ln(N_{hi1}) - \ln(N_{hi} P_{hi1})$ , which is the natural log of observed bin-level correct responses less the natural log of expected bin-level correct responses. An example of the likelihood ratio type statistic is the  $G^2$  statistic (McKinley & Mills, 1985).

The differentiating properties of the various indices and tests of model-data fit are based on three primary dimensions. The first dimension is whether the chi-square or likelihood ratio approach is taken. The next dimension is the manner in which bins are defined. The bins can be defined in several different ways, from having each individual serving as a unique bin, to having bins defined according to a particular number of individuals. The final dimension is the manner in which  $P_{hi1}$  is computed.

One common criticism of traditional chi-square type approaches (e.g., Yen's  $Q_I$ ) and traditional likelihood ratio approaches (e.g.,  $G^2$ ) is their use of ability estimates ( $\hat{\theta}$ ) for creating bins of people. As Yen (1981) discusses, a poorly fitting model could result in biased ability estimates. Thus, binning on biased ability estimates may provide an invalid item-fit statistic. This limitation highlights the need for alternative approaches for creating bins that are not based on  $\theta$ .

One approach for creating bins that are  $\theta$ -independent is to define bins according to observed test scores (e.g., summated scores) rather than  $\theta$  estimates (Orlando & Thissen, 2000). For instance, if an assessment has 15 items, there will be 14 bins representing those earning a summated score of 1, 2, ..., 12, 13, 14 on the assessment. The bins range from 1 to  $I - 1$  because the probability of correct response for an item is always zero at a summated score of 0 (in which a person answers no items correct) and is always 1 at a summated score of  $I$  (in which a person answers all items correct). Applying this approach to the chi-square form of Equation 13 yields the  $S-X^2$  statistic, and applying this approach to the likelihood-ratio approach of Equation 14 yields the  $S-G^2$  statistic (Orlando & Thissen, 2000). For both  $S-X^2$  and  $S-G^2$ ,  $df = I-1$  less the number

of model parameters estimated by the model. For instance, for a 15-item test, with the item of interest estimated using a 3-PL,  $df = 15 - 1 - 3 = 11$ .

Orlando and Thissen (2000) argued that because the expected proportion of correct responses from the IRF is based on model-dependent ability estimates, the statistic's distribution is unclear and conclusions drawn from these statistics may be invalid. To address this issue,  $S-X^2$  and  $S-G^2$  compute  $P_{hi1}$  using a recursive algorithm. The reader is referred to Lord and Wingersky (1984) for a description of this recursive algorithm.

Two related fit indices that follow the chi-square approach of Equation 13 are *OUTFIT* and *INFIT* (Wright & Panchapakesan, 1969). While these approaches are based on the same general form of Yen's  $Q_i$ , they adopt a notably different approach to bin definition. Both *OUTFIT* and *INFIT* assign only one individual per bin, such that each individual serves as a unique bin. Therefore, the residual adopted in Equation 13 for the chi-square approach is the individual-level residual. In addition, because there is a single individual per bin, the observed proportion correct for each bin is simply the individual's scored response to the item (i.e., 0 or 1), and the value of  $P_{hi1}$  is the value of the item's IRF for correct response at the individual's estimated ability level.

While *OUTFIT* and *INFIT* share the property of having one individual per bin, and thus adopting the individual-level residual of Equation 2, they differ in how much weight they assign to each individual. *OUTFIT* is computed using

$$OUTFIT_i = \frac{1}{N} \chi_i^2, \quad (15)$$

where  $\chi_i^2$  is the general chi-square form represented in Equation 4 with one individual per bin ( $N_{hi} = 1$  for all  $h$ ) and  $N$  represents the total number of individuals in the sample. Because *OUTFIT* divides  $\chi_i^2$  by  $N$ , *OUTFIT* is not actually a chi-square statistic, but rather an index of the magnitude of lack of fit that can be interpreted as the typical squared, standardized residual in the sample. Values of *OUTFIT* close to 1 indicate good model-data fit and values much greater, or much less, than 1 indicate the model-data fit is problematic. One suggestion has been to flag an item as misfitting if *OUTFIT* is less than 0.5 or greater than 1.5 (de Ayala, 2009, pg. 53) or use a transformed  $t$  statistic with values less than -2 or greater than 2 indicating misfit (Bond & Fox, 2007; de Ayala, 2009).

*OUTFIT* assigns each individual the same weight in its computation, which can be a limitation because it can be heavily impacted by the potential of very large individual-level residuals of people for which the probability of correct response is considerably low or considerably high. *INFIT* addresses this limitation by assigning more weight to individuals having an ability level ( $\theta$ ) closer to the item difficulty value ( $b_i$ ). An individual whose ability is close to the item's difficulty should give better insight into that item's performance than an individual who has ability that is substantially different than item difficulty. The weight assigned to each individual-level residual is equal to the Rasch model information function (see Hambleton et al. 1991 for an accessible description of the information function) at the individual's level of ability, which is given by  $P_{hi1}(1 - P_{hi1})$ . This leads to *INFIT* being less sensitive to extreme responses than *OUTFIT* (Bond & Fox, 2007). For this reason, stronger consideration typically is given to *INFIT* (de Ayala, 2009). The general heuristic for flagging an item as misfitting using

*INFIT* is similar to that used for *OUTFIT*; items are flagged when *INFIT* values are less than 0.5 or greater than 1.5. Similarly, the transformed *t* statistic is also provided for *INFIT* in Winsteps (Linacre, 2011), with values less than -2 or greater than 2 indicating poor model-data fit.

### **Bayesian Approaches to Model-Data Fit**

There are several options, which are similar to the Frequentist approaches, for Bayesian evaluation of model-data fit. These include (a) examining the sensitivity of inferences to changes in the prior distribution and the analysis model, (b) checking the sensibility of posterior inferences against the researcher's substantive knowledge, and (c) checking that the model fits the data (Gelman, Meng, & Stern, 2003). The majority of the focus of Bayesian IRT model-data fit has come via the latter method, typically assessed via PPC (Gelman, Meng, & Stern, 2003). The PPC method consists of assessing the plausibility of PPD against observed data, similar to the concept of residuals in the Frequentist paradigm discussed above. In both approaches, if the predictions resemble the observations, then the model is deemed a good fit to the data. If the model is a good fit to the data, then future data simulated from the model should look very much like observed data. 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 PPC is as follows. First, PPD data is simulated from the parameter posteriors. The data is simulated by randomly drawing values from the posterior distributions of item and person parameters and generating data which would likely arise if these were true parameter values. Another set of values is then drawn

randomly and used to generate another data set which would likely arise from this second value. This process is repeated until the desired number of simulated data sets is generated. Next, a comparison is made between the simulated and observed data. If the data sets are similar, the conclusion is that the model fits the data well (Lynch, 2007). A diagram of the PPC approach is found in Figure 3.

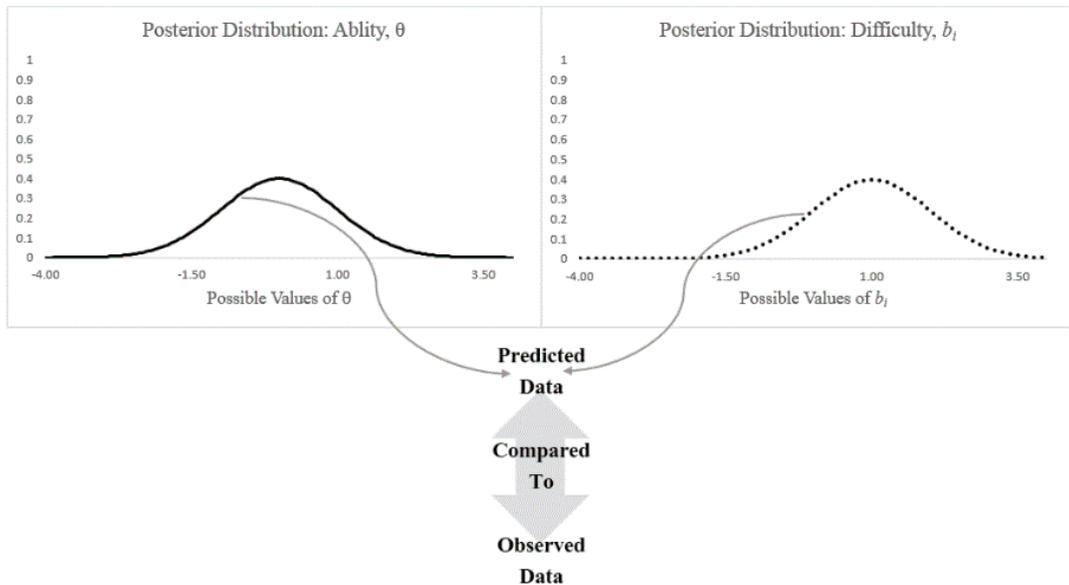


Figure 3. Illustration of Posterior Predictive Checks

Tests using Bayesian p-values are available for drawing conclusions regarding the similarity of the simulated and observed data sets. Let  $T(x)$  be a statistic applied to the observed data, where the observed data is denoted by  $x$ . The statistic  $T(\cdot)$  could be any commonly available item fit statistic, such as  $G^2$  or  $S-X^2$ , or other descriptors of the data, such as the observed percent correct on an item. 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  $p$ -value is

$$p - value = proportion(T(x^{sim}) \geq T(x)). \quad (16)$$

This posterior predictive  $p$ -value (PPP) 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 (Sinharay, 2006). One complication of using the PPP value is that it can be sensitive to small samples (Meng, 1994).

There is no limit on the number of statistics that could be used to obtain Bayesian PPP values, illustrating the flexible nature of the Bayesian method (Lynch, 2007). The term *discrepancy measure* is defined as the use of the discrepancy, or difference, between observed and simulated data in the PPC analysis (Meng, 1994). The previous discussion of Frequentist approaches illustrates possible discrepancy measures for the PPC approach. However, careful consideration should be given to the choice of discrepancy measure. For instance, Sinharay and Johnson (2003) found that the use of the percentage correct (or percent of respondents per category) as a discrepancy measure did not permit model-data misfit detection. Toribio and Albert (2011) investigated *OUTFIT* (Wright & Panchapakesan, 1969), Yen's  $Q_I$  (Yen, 1981),  $G^2$  (McKinley & Mills, 1985),  $S-X^2$  and  $S-G^2$  (Orlando & Thissen, 2000), finding all performed equally well as discrepancy measures. However, their computation of  $S-X^2$  and  $S-G^2$  was a simplified version of the

original statistics' intended methodology and their results might not be applicable for these discrepancy statistics.

To summarize the discussion thus far, IRT models have a set of assumptions that must be checked for adequate specification before valid inferences can be drawn. One such assumption is that the form of the IRF is appropriate for the data at hand.

Frequentist methods typically apply one of two approaches, either a chi-squared statistic or a likelihood ratio statistic. In the Bayesian framework, evaluation of model-data misfit is often assessed through the PPC method. There has been considerable research on the effects of model-data misfit on parameter estimation in the Frequentist framework.

However, the effect of model-data misfit on posterior estimation has been given relatively little attention.

### **Prior Specification in PPC**

PPC has proven to be a useful tool (Sinharay, 2006), but there has been little consistency in the specification of priors in this area. Multiple studies (Sinharay et al., 2006; Sinharay, 2005; Sinharay & Johnson, 2003) have used noninformative priors for dichotomous models such as  $\log(a_i) \sim \text{Normal}(0, 10)$  and  $b_i \sim \text{Normal}(0, 10)$ .

Similarly, Ames (2014) used noninformative priors for PPC applied to polytomous models (truncated  $\log(a_i) \sim \text{Normal}(0, 10)$  and  $b_i \sim \text{Normal}(0, 10)$ ). In contrast, Zhu and Stone (2011) used more narrowly focused priors for the GRM such as  $\log(a_i) \sim \text{Normal}(0, 1)$  and  $b_{ik} \sim \text{Normal}(0, 1)$ . Sinharay (2006) also used informative priors. Another study on dichotomous items (Toribio & Albert, 2011) failed to provide any information on the prior parameters used.

Parameter recovery studies involving Bayesian methods have also varied the use of priors. Sheng (2010) varied the informativeness of the parameter priors for  $a_i$  and  $b_i$  as follows: (a) noninformative uniform prior, (b) noninformative normal prior with a large variance ( $10^{10}$ ) and mean of 0, (c) more informative prior (with a variance of 4 and mean of 0), and (d) precise prior (with a variance of 1 and mean of 0). Three priors were considered for guessing: (a) noninformative Beta prior ( $Beta(1,1)$ ), (b) informative Beta prior with mean 0.22 and standard deviation of 0.131 ( $Beta(2,7)$ ), and (c) very informative Beta prior with mean 0.22 with smaller standard deviation of 0.007 ( $Beta(5,17)$ ). Sheng found that relatively informative priors, when accurately specified, should be adopted for the discrimination and difficulty parameters.

Kieftenbeld and Natesan (2012) used noninformative priors for parameter recovery of the GRM including log-normal priors for the discrimination parameters ( $\log(a_i) \sim N(0, 2^{1/2})$ , implying  $\text{mean}(a_i) = 2.718$  and  $\text{var}(a_i) = 6.871$ ) and a noninformative uniform prior on the interval  $(-5, 5)$  for the threshold parameters (subject to an ordering restriction). They found that almost all discrimination parameters were positively biased, likely due to the mean of the noninformative prior being larger than any of the generating discrimination values. In Wollack, Bolt, Cohen, and Lee (2002), all priors were distributed  $\text{Normal}(0, 100)$  so as to have a “negligible effect on item parameter estimation” of the nominal response model.

The above-mentioned studies were based upon recovery of parameters with good model-data fit. However, despite the good model-data fit, Kieftenbeld and Natesan (2012) found that the noninformative priors likely induced positive bias in some item

parameter estimates when using the posterior mean as a point estimate. This bias is potentially problematic for application of the PPC method because data are simulated from draws from the posteriors. If data are simulated from posteriors that are the result of model misspecification, then the simulated data will likely look different from observed data. The PPC method will then find evidence of misfit due to the misspecified posterior and not the actual presence of misfit. It is clear from this that the use of noninformative priors in the PPC method requires more attention.

With model-data misfit, the likelihood tends to be flatter than those likelihoods arising from good model-data fit (Drasgow et al., 1987). Drasgow and colleagues (1987) examined nine appropriateness measures to identify inappropriate test scores. In each approach, response patterns were quantified to determine the degree to which an observed response vector is atypical, an approach similar to determining aberrant responses for an item across individuals. In their paper, they discussed two indices, the jackknife variance estimate (Mosteller & Tukey, 1968) and a comparison of the expected and observed likelihood curvatures (Efron & Hinkley, 1978), which provide a measure of the flatness of a likelihood function. These indices are based upon the notion that responses not fitting the expected or predicted pattern will flatten the likelihood function near its maximum because no single ability parameter estimate exists that will provide a good fit to the response profile. The result is that the likelihood is relatively flat and will not have a sharp maximum (Drasgow et al., 1987).

The flatter the likelihood, the larger the role of the prior in formation of the posterior distribution (see again Figure 1; Muller, 1989). Thus, specification of the prior

in such studies requires careful consideration. With both a flat prior and flat likelihood, the posterior will tend to be flat as well, providing very little information and possibly affecting the results of the PPC method. The result would be potentially assigning large probabilities to a very broad range of parameter values.

Simulating data from this type of posterior would provide very inconsistent predictive data, since parameter values would be chosen from a broad range of values, and the resulting data sets might or might not differ in a systematic way. As an example, assume a draw is being taken from a parameter posterior which resembles a standard normal distribution ( $\text{Normal}(0, 1)$ ). If the true parameter value is actually centered one standard deviation above the estimated parameter posterior (e.g.,  $\text{Normal}(1, 1)$ ), then predictive data will likely be systematically different than observed data. However, this may depend on the values chosen for the flat parameters. For example, for a parameter posterior coming from a flat prior and a flat likelihood ( $\text{Normal}(0, 10)$ ), the predictive data has relatively the same chance for appearing similar to the observed data as it has for appearing different from it.

Checking the Bayesian model also implies checking for prior-data conflicts – situations in which the prior's distribution does not match that of the observed data (Evans & Moshonov, 2006) - but the PPC method only addresses sampling model appropriateness. Evans and Moshonov (2006) claim that if, when checking the appropriateness of the sampling model, the researcher finds that the model is inappropriate, checking for prior-data conflict isn't necessary. However, they provide no evidence that prior-data conflicts do not confound model checking procedures such as

PPC. It would seem in the case of Kieftenbeld and Natesan (2012) that the misspecification of the prior may have induced model misfit which would have been detected by the PPC method. Clearly the prior-data conflict in Kieftenbeld and Natesan (2012) would have interfered with the PPC method if the researchers had gone on to investigate model-data fit.

Two attempts to circumvent the prior-data conflict have been used in simulation studies investigating the PPC method. The first is to use large sample sizes when generating the data sets, in the hope is that prior will have less of a role in MCMC parameter estimation and, therefore, the PPC method (Sinharay & Johnson, 2003). The large sample size also has been used to ensure that the analysis model is estimated precisely and that the PPC results will not be affected by any inaccuracy in model parameter estimation (Zhu & Stone, 2011). For instance, Sinharay and Johnson (2003), Sinharay (2006), and Sinharay et al. (2006) used 2500 individuals. Some smaller sample sizes were investigated by Toribio and Albert (2011), who used 1000 individuals. However, this procedure undermines one of the benefits of Bayesian methods – that they are desirable with sparse data and small sample sizes when asymptotic theory is unlikely to hold and Frequentist approaches are limited (Fox, 2010). Further, parameter estimates from MCMC estimation are influenced not only by the sample size, but by the specification of the prior distribution.

The second approach employed to avoid prior-data conflict has been to specify noninformative priors when conducting the PPC method. As previously discussed, noninformative priors are often used so that the data drives the MCMC estimation

process rather than the priors (Sinharay, 2006). However, as previously discussed, the use of noninformative priors may have problems when used with the PPC method.

Gelman, Bois, and Jiang (1996) commented that if parameters are well-estimated from the data, PPC give results similar to classical model-checking procedures, irrespective of the “reasonable” prior used. But the parameters may be poorly estimated in the presence of model-data misfit and prior specification may become more important. For instance, with model-data misfit, there are known problems with parameter estimation, some of the most critical problems being IRT parameter invariance no longer holds and bias in the ability estimates (Bolt, 2002; Rupp & Zumbo, 2004; Shepard, Camilli & Williams, 1984). The biased ability estimates are potentially problematic for the PPC method. In the data simulation step of the PPC, draws are taken off the posteriors of both item and ability parameters. If either, or both, of the posteriors are biased, the resulting PPD sets might also be biased. In this case, the PPC method would also be checking the accuracy of posterior estimates and not just model-data misfit. Stern and Sinharay (2005) extended Gelman and colleague’s concern about the effect of prior distributions on PPC methods. They stated that model failures are detected only if the posterior inferences under the likelihood model seem flawed. If the prior used is unsuitable, posterior inferences may still be deemed reasonable if the prior has little effect on the posterior, such as the case with a flat prior or large sample sizes (p. 178). However, again, this argument seems to rely upon the flatness of the prior relative to the likelihood.

Two more studies address the effect of priors on the PPC method. In the first, Gelman, Meng and Stern (1996) apply PPC to a study fitting a latent two-class mixture model to the data from an infant temperament study. Dirichlet parameter priors were chosen so that the multinomial probabilities for a variable (e.g., motor activity) were centered on values elicited from psychological theory, but with a large variance. The use of a weak, but not uniform, prior distribution was used to help identify the mixture classes (Gelman et al., 1996). The authors computed PPP values under a variety of prior distributions. The center of each class of the prior was chosen to either match the values suggested by theory or to represent a uniform distribution. The strength of the prior (informativeness) was also varied.

Gelman and colleagues found that as long as the prior distributions were not particularly strong, the center of the prior distribution had little effect on the PPC method and the size of the PPP values remained relatively constant (Gelman et al., 1996). With incorrect and very strong priors (which are the opposite of noninformative in that they narrow the mass of the likelihood onto a narrow region), the PPP value could be quite misleading. However, with correct and very strong priors, the PPP values reflected true model-data misfit rates of the mixture models. Sinharay and Johnson (2003) note (in connection with Gelman et al.'s findings) that a strong prior distribution, with reasonable trustworthiness, can be used to more effectively assess the fit of the likelihood part of the model. However, despite this recommendation, Sinharay and Johnson (2003) used large samples and noninformative priors for their PPC study.

Often, informative priors may prove more useful to IRT practitioners than noninformative priors. For instance, Fox (2010) concluded that the elicited hierarchical prior proved more useful for the 2PL model than did noninformative priors, especially with relatively small datasets when prior information can significantly influence the item parameter estimates. Tsutakawa (1992) used information from a previous years' test administration to help guide the specification of elicited priors. When comparing joint maximum likelihood to Bayesian estimation, Gifford and Swaminthan (1990) found specification of the priors had modest effects on the Bayesian estimates, but concluded that the effect of the prior was greater for more complex models, particularly for the lower asymptote parameter of the 3PL model. Similarly, Swaminathan and colleagues (2003) found that the incorporation of ratings provided by subject matter experts produced estimates that were more accurate than those obtained without using such information. The improvement was observed for all item response models, but the improvement was positively related to the number of parameters estimated. Thus, as model complexity grew, the need for specifying informative priors grew in importance.

Albert and Ghosh (2010) showed that noninformative priors for IRT models often lead to improper posteriors. Because of this, their recommendation is to choose a prior that is proper to ensure that posterior distributions are also proper. They state this is most important with extreme data, which is often the type of data simulated in IRT model-data fit studies, such as outliers and unexpected response patterns, which would then be investigated via the PPC method.

A final point related to the prior's influence on PPC methods concerns the manner by which the Bayes factor summarizes the evidence provided by the data in favor of one statistical model in comparison to another. (Please see Kass and Raftery (1995) for a comprehensive review of Bayes factors, including information about their interpretation.) There has been considerable research in the area of the role of the prior in the use of the Bayes factor, which is known to be highly sensitive to choice of the prior distribution (Sinharay & Stern, 2002). However, the role of the prior with PPC has not received the same attention as the Bayes factor. Therefore, the need for investigating the role of the prior on the PPC method warrants the more attention.

In summary, because IRT model-data misfit results in a flatter likelihood, use of noninformative priors with the PPC method could be problematic. It might be the case that using more informative priors might prove more useful for conducting tests of model-data misfit than using noninformative priors. This is because the likelihood will have less of an effect on the posterior with increasing model-data misfit.

## **CHAPTER III**

### **METHODS**

This chapter presents the methodology which is proposed to address the gap identified in the literature on prior specification for the PPC method. To summarize the literature found in the previous chapter, the PPC method might be sensitive to choices in specification of the prior distributions because of the flatness of the likelihood in the presence of model-data misfit. The likelihood will show decreased curvature as model-data misfit increases. With increased flatness, the prior will have more of an effect on posterior distributions and PPD.

Consistency of the studies in PPC has been lacking in regards to how priors have been specified. However, it is common for researchers to use noninformative priors in this approach, relying on the belief that the noninformative prior will have little effect on the procedure (Sinharay, 2006). This may be true for models exhibiting adequate model data fit and a sufficiently peaked likelihood. However, with model-data misfit present, the likelihood will be less peaked, and the posterior may be more influenced by the noninformative prior than initially intended.

A posterior resembling a noninformative prior would assign high probabilities to parameters across a very broad range of values. Thus, when sampling from the posterior to simulate PPD, a wide range of parameter values would have a nearly identical probability of being sampled and systematic differences may be more difficult to detect.

Further, IRT parameters have specific distributional assumptions and noninformative priors may not be appropriate in the theoretical realm either.

The methods discussed in this chapter will provide guidance on assessing the PPC approach's sensitivity to prior specification as well as under which conditions researchers must be most attentive to the choice of priors for PPC. The four research questions posited at the end of Chapter 1 will guide the research design and methodology described in this section.

### **Research Plan**

This section describes the research plan, including simulation conditions and other considerations.

### **Choice of Simulation Conditions**

**General.** Harwell and colleagues (1996) recommend no fewer than 25 replications for simulation studies in IRT. With improved computational efficiency (see Ames & Samonte, in press), studies involving MCMC methods can now perform more than 25 replications. Levy et al. (2009) used 50 replications. Sinharay (2006), Sinharay and Johnson (2003), Sinharay et al. (2006) and Toribio and Albert (2011) all used 100 replications. Further, Sinharay subsequently has recommended a minimum of 100 replications (personal communication, April 14, 2014). To be in keeping with previous studies, 100 unidimensional, dichotomous IRT data sets for each condition will be simulated, with each data set representing one replication for the simulation.

This study will use the methodological approach described in Gelman, Meng, and Stern (1996) to explore the role of prior specification on the PPC method and the effects

of sample size and type of misfit on the role of the prior, on the PPC method. This procedure was also followed in Sinharay (2006), Sinharay and Johnson (2003), and Sinharay et al. (2006).

Consider a data-generating model (GM) and an analysis model (AM), where the AM is never more complex than the GM. Each GM model may be one chosen from the 2PL, 3PL, and logistic positive exponent (LPE; Samejima, 2000) models. When the GM is the LPE, the AM include the LPE, 3PL, 2PL, and 1PL. When the GM is the 3PL, the AM include the 3PL, 2PL, and 1PL. When the GM is a 2PL model, the 2PL and 1PL could be used as AM. Situations in which the GM is equal to the AM, good model-data fit is expected. However, when the GM is not the same as the AM, model-data misfit may occur. An explanation of the introduction of model-data misfit follows in the next section. Figure 4 provides an illustration of the possible GM-AM combinations.

		Analysis Model			
		1PL	2PL	3PL	LPE
Generating Model	2PL	X	X		
	3PL	X	X	X	
	LPE	X	X	X	X

Figure 4. Generating and Analysis Model Combinations

**Introduction of Misfit.** Simulating misfit in IRT models should be given careful consideration because the type of IRT misfit may be one of several, depending in part on the model being fit (Wells & Bolt, 2008). Wells and Bolt (2008) consider several sources of misfit, distinguished by the location along the latent ability scale at which the largest amount of misfit was introduced. Several types of misfit are illustrated in Figure 5.

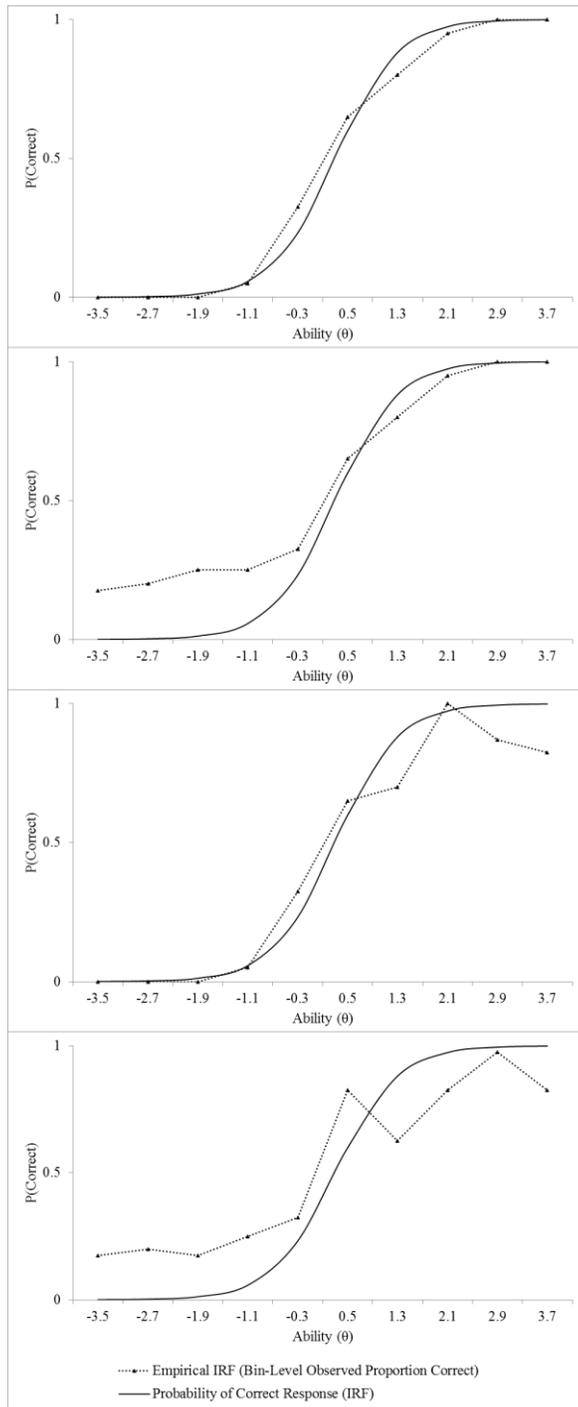


Figure 5. Types of Misfit in IRFs

Good model-data fit is expected when the GM is the same type as the AM. This is illustrated in the top panel of Figure 5. However, when the AM is a less complex version of the GM, model-data misfit can arise. For instance, consider the case when the GM is the 3PL with a lower asymptote of  $c_i = 0.2$ . If a restricted AM, such as the 2PL, is fit to the data simulated from the 3PL GM, misfit will be present in the lower range of  $\theta$ . This type of misfit is evident in the second panel of Figure 5. There are several consequences of such a type of misfit, such as difference in difficulty parameter estimates between the two models. The difficulty estimates will be, in general, lower for the 2PL than for the 3PL, resulting in the appearance that the items are less difficult than the true item difficulty (Bergan, 2010). Another consequence of this type of misfit is that the test information curves can be misleading, with the correctly specified 3PL showing a test information function that reaches its maximum at a value of  $\theta$  higher than that of a test information function for a (effectively) misfitting 2PL model.

Most attention related to the specification of the functional form of the IRF has been in the context of fitting a traditional AM to a traditional GM, such as a combination chosen from among the 3PL, 2PL, or 1PL. Another type of misfit is introduced when the GM is the LPE and the AM is a simpler, traditional IRT model. Bolt and colleagues (2014) introduced this type of misfit in the context of model misspecification for measuring growth in vertical scaling. A LPE-2PL model is represented by the following equation:

$$P_{ni}(x_{ni} = 1 | \theta_n, a_i, b_i, \xi_i) = \left[ \frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))} \right]^{\xi_i}, \quad (17)$$

where  $\xi_i$  is the acceleration parameter, representing the complexity of the item. In reality, any of the 1PL, 2PL, or 3PL models in Equations 5-7 could serve as the basis for the LPE, with appropriate restrictions, similar to restrictions imposed on traditional IRT models. The 3PL version of the LPE (LPE-3PL) is represented via

$$P_{ni}(x_{ni} = 1 | \theta_n, a_i, b_i, c_i, \xi_i) = c_i + (1 - c_i) \left[ \frac{\exp(a_i(\theta_n - b_i))}{1 + \exp(a_i(\theta_n - b_i))} \right]^{\xi_i}. \quad (18)$$

The acceleration parameter in Equations 17 and 18 introduces asymmetry to the IRF. This is accomplished through accelerating (i.e., pushing higher) the ability location at which the IRF's slope is maximized. When  $\xi_i = 1$ , the IRFs are symmetric and the formula reduces to the traditional form of the IRF. When  $\xi_i > 1$ , the asymmetry is such that the IRF more rapidly increases on the left side of the inflection point than it decelerates on the right side of the inflection point (Bolt et al., 2014; Samejima, 2000). With  $\xi_i < 1$ , just the opposite occurs and the LPE results in asymmetric ICCs with positively skewed slopes. One implication of this is that the estimated discrimination parameter of data fitted to a traditional 2PL, or 3PL, will generally be lower when estimated for a group of higher ability. The role of the acceleration parameter on the shape of the IRF is illustrated for the 2PL version of the LPE in Figure 6. The amount of misfit generated through the LPE in relation to traditional IRFs is not substantial, but can be used to introduce misfit at varying locations along the  $\theta$  scale. One example is found in Figure 5 in the third panel, where misfit occurs at the higher range of  $\theta$ .

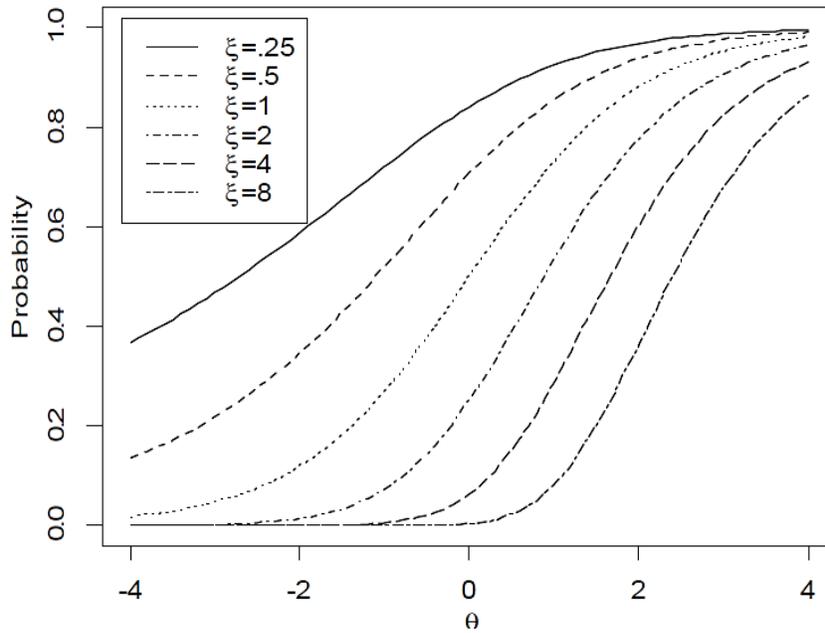


Figure 6. IRFs for Hypothetical 2PL LPE Items ( $a_i = 1$ ,  $b_i = 0$  for all items).

The magnitude of the misfit is equally as important as the location (Bolt, 2002; Wells & Bolt, 2008). One option for quantifying the magnitude of misfit is found by summing the weighted differences between the GM and AM at  $k$  equally spaced  $\theta$  values, ranging from -3 to 3 as follows

$$MISFIT = \sqrt{\sum_{k=1}^{601} w(\theta_k)(P_{GM,k} - P_{AM,k})^2} , \quad (19)$$

where  $w(\theta_k)$  is a normalized weight defined by the standard normal density. The probability of correct response at the  $k^{th}$  level of  $\theta$ , according to the nonparametric curve of the generating model (i.e., the empirical IRF), is denoted by  $P_{GM,k}$  and  $P_{AM,k}$  is the IRF of the Bayesian model fit to the data using the posterior Bayes modal estimate (Wells &

Bolt, 2008). The weighting factor is used to weight the difference between the empirical and modeled curve according to the expected number of examinees at each ability location.

This study will simulate two classes of misfit. Following Wells and Bolt (2008), a cut-off of 0.020 will be used to distinguish between items with small versus medium to large misfit. If an item exhibits a MISFIT value of 0.020 or above, it will be considered as a medium to large misfit. If an item exhibits a MISFIT value of below 0.020, it will be considered as a small misfit.

**Data-generating parameters.** In keeping with the approach of Sinharay, Johnson, and Stern (2006), test length will be fixed at 30 items. In their study, when the GM was the 3PL, all of the item parameters found in Table 2 were used. When the GM was the 2PL, only the discrimination and difficulty parameters were used and when the 1PL was used, the average of the item discriminations ( $a=1.36$ ) and difficulty parameters was used. This approach appears to be a common practice in PPC simulation studies, with the technique adopted again in Sinharay and Johnson (2003), Sinharay (2006), and Toribio and Albert (2011).

Table 2. Generating Model Parameters in Sinharay, Johnson, and Stern (2006)

<b>Item ID</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>
$a_i$	0.70	1.89	1.32	1.36	1.17	0.56	1.10	1.68	1.01	0.88
$b_i$	1.81	-0.52	0.26	-1.48	-0.52	0.44	-2.15	0.96	-0.87	1.70
$c_i$	0.08	0.06	0.14	0.17	0.17	0.14	0.30	0.25	0.10	0.22
<b>Item ID</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>	<b>17</b>	<b>18</b>	<b>19</b>	<b>20</b>
$a_i$	1.19	0.60	1.49	2.01	2.40	2.00	1.48	1.10	1.52	1.45
$b_i$	-2.41	-0.56	-0.27	0.04	-0.79	-0.38	0.14	-1.53	0.11	-0.21
$c_i$	0.23	0.21	0.12	0.20	0.18	0.17	0.13	0.14	0.20	0.12
<b>Item ID</b>	<b>21</b>	<b>22</b>	<b>23</b>	<b>24</b>	<b>25</b>	<b>26</b>	<b>27</b>	<b>28</b>	<b>29</b>	<b>30</b>
$a_i$	0.80	0.67	0.83	1.17	1.43	2.40	1.53	1.20	1.70	2.05
$b_i$	-0.52	-0.43	-1.25	-0.52	-0.27	-0.44	1.75	-0.80	0.40	-0.93
$c_i$	0.11	0.12	0.18	0.13	0.15	0.40	0.25	0.24	0.16	0.26

However, with the item parameters in Table 2 and the approach described above, potential problems can occur. To illustrate, consider a scenario with 2500 individuals randomly drawn from a standard normal distribution. Let the GM be the 3PL and the AM be the 2PL. The MISFIT values for the 30 items under this illustration are found in Table 3, showing that 19 of the 30 of the items (63%) had MISFIT values classified as medium to large. The items with the smallest MISFIT values were those with difficulty parameters below 0, indicating that the 2PL tends to adequately fit a 3PL item as long as the item is relatively easy, a finding consistent with Wells and Bolt (2008). In Table 3, an

asterisk indicates the item has been classified as having medium to large misfit based on a cutoff of 0.02.

Table 3. MISFIT for GM=3PL and AM=2PL

Item	MISFIT	Item	MISFIT
1	0.017	16*	0.030
2	0.011	17*	0.028
3*	0.024	18	0.009
4*	0.021	19*	0.032
5*	0.021	20*	0.031
6*	0.020	21	0.014
7	0.009	22	0.008
8*	0.057	23	0.016
9	0.008	24*	0.028
10*	0.025	25*	0.020
11	0.003	26*	0.055
12	0.019	27*	0.050
13*	0.026	28*	0.021
14*	0.044	29*	0.040
15*	0.031	30*	0.026

The preponderance of misfitting items has implications for this study. When multiple misfitting items are present, dependency in the items can occur (Yen, 1981). This dependency results in parameter estimation of one item which is not independent from the parameter estimation of another item. Therefore, having 63% of the items simulated with medium to large misfit could confounded parameter recovery with misfit. In an effort to control for this confounding, this study will implement varying levels of the degree of representation of misfitting items on the test. Following the approach outlined by Wells and Bolt (2008), there will be four levels regarding the percent of

misfitting items: 0% ( $n_{items} = 0$ ), 10% ( $n_{items} = 3$ ), 30% ( $n_{items} = 9$ ), and 50% ( $n_{items} = 15$ ). The generating item parameters for the study can be found in Tables 4 through 6, with detail provided below.

For the condition where the GM is the LPE-2PL, the generating item parameters are found in Table 4. When the GM is the LPE, the AM include the LPE, 3PL, 2PL, and 1PL. Parameters were borrowed from Table 2 (originally found in Sinharay et al., 2006) when items were intended to induce misfit. The 2PL-LPE (Equation 17) was chosen rather than the 3PL-LPE (Equation 18) to isolate the location of the misfit to the higher end of the ability scale only, as the 3PL GM can induce misfit at the lower end of the ability scale. To further induce this form of misfit, only acceleration parameters greater than 1 were considered.

To ensure that the items introduce adequate levels of misfit, a MISFIT index will be computed for each of the items in each replication of the conditions. An approximate value of the MISFIT index, using quadrature rather than simulated data, can be found for the GM-LPE, AM combinations in Table 4 using quadrature.

For the condition where the GM is the 3PL, the generating item parameters are found in Table 5. When the GM is the 3PL, the AM include the 3PL, 2PL, and 1PL. An approximate value of the MISFIT index, using quadrature, can be found for the GM-3PL, AM combinations in Table 5. For the condition where the GM is the 2PL, the generating item parameters are found in Table 6. When the GM is the 2PL, the AM include the 2PL and 1PL. An approximate value of the MISFIT index, using quadrature, can be found for

the GM-2PL, AM combinations in Table 6. In Tables 4-6 it can be seen that MISFIT values tend to be larger when the GM is the LPE or 3PL than when the GM is the 2PL.

Table 4. LPE-2PL Generating Model Parameters

I	0% Misfitting			10% Misfitting			30% Misfitting			50% Misfitting			MISFIT		
	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	AM- 3PL	AM-2PL	AM-1PL
1	1	-2	1	1	-2	1	1	-2	1	1	-2	1	--	--	--
2	1	-1.9	1	1	-1.9	1	<b>1.45</b>	<b>-0.21</b>	<b>1.5</b>	<b>1.45</b>	<b>-0.21</b>	<b>1.5</b>	<b>0.022</b>	<b>0.015</b>	<b>0.039</b>
3	1	-1.7	1	1	-1.7	1	1	-1.7	1	<b>1.17</b>	<b>-0.52</b>	<b>2</b>	<b>0.066</b>	<b>0.062</b>	<b>0.060</b>
4	1	-1.6	1	1	-1.6	1	1	-1.6	1	1	-1.6	1	--	--	--
5	1	-1.4	1	1	-1.4	1	<b>1.53</b>	<b>1.75</b>	<b>2</b>	<b>1.53</b>	<b>1.75</b>	<b>2</b>	<b>0.084</b>	<b>0.035</b>	<b>0.053</b>
6	1	-1.3	1	1	-1.3	1	1	-1.3	1	<b>0.56</b>	<b>0.44</b>	<b>4</b>	<b>0.157</b>	<b>0.123</b>	<b>0.120</b>
7	1	-1.2	1	1	-1.2	1	1	-1.2	1	1	-1.2	1	--	--	--
8	1	-1.0	1	1	-1.0	1	1	-1.0	1	<b>1.48</b>	<b>0.14</b>	<b>2</b>	<b>0.093</b>	<b>0.060</b>	<b>0.065</b>
9	1	-0.9	1	1	-0.9	1	1	-0.9	1	1	-0.9	1	--	--	--
10	1	-0.8	1	<b>1.68</b>	<b>0.96</b>	<b>2</b>	<b>1.68</b>	<b>0.96</b>	<b>2</b>	<b>1.68</b>	<b>0.96</b>	<b>2</b>	<b>0.091</b>	<b>0.049</b>	<b>0.066</b>
11	1	-0.6	1	1	-0.6	1	1	-0.6	1	<b>1.32</b>	<b>0.26</b>	<b>4</b>	<b>0.140</b>	<b>0.108</b>	<b>0.110</b>
12	1	-0.5	1	1	-0.5	1	1	-0.5	1	1	-0.5	1	--	--	--
13	1	-0.3	1	1	-0.3	1	<b>2.40</b>	<b>-0.79</b>	<b>4</b>	<b>2.40</b>	<b>-0.79</b>	<b>4</b>	<b>0.105</b>	<b>0.087</b>	<b>0.077</b>
14	1	-0.2	1	1	-0.2	1	1	-0.2	1	1	-0.2	1	--	--	--
15	1	-0.1	1	1	-0.1	1	<b>2.01</b>	<b>0.04</b>	<b>1.5</b>	<b>2.01</b>	<b>0.04</b>	<b>1.5</b>	<b>0.067</b>	<b>0.010</b>	<b>0.053</b>
16	1	0.1	1	1	0.1	1	1	0.1	1	1	0.1	1	--	--	--
17	1	0.2	1	1	0.2	1	1	0.2	1	1	0.2	1	--	--	--
18	1	0.3	1	1	0.3	1	1	0.3	1	<b>1.53</b>	<b>1.75</b>	<b>2</b>	<b>0.084</b>	<b>0.035</b>	<b>0.054</b>
19	1	0.5	1	1	0.5	1	1	0.5	1	1	0.5	1	--	--	--
20	1	0.6	1	<b>1.70</b>	<b>0.40</b>	<b>4</b>	<b>1.70</b>	<b>0.40</b>	<b>4</b>	<b>1.70</b>	<b>0.40</b>	<b>4</b>	<b>0.129</b>	<b>0.097</b>	<b>0.101</b>
21	1	0.8	1	1	0.8	1	1	0.8	1	1	0.8	1	--	--	--
22	1	0.9	1	1	0.9	1	1	0.9	1	1	0.9	1	--	--	--

I	0% Misfitting			10% Misfitting			30% Misfitting			50% Misfitting			MISFIT		
	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	$a_i$	$b_i$	$\xi_i$	AM- 3PL	AM-2PL	AM-1PL
23	1	1.0	1	1	1.0	1	1	1.0	1	<b>0.60</b>	<b>-0.56</b>	<b>2</b>	<b>0.099</b>	<b>0.072</b>	<b>0.085</b>
24	1	1.2	1	1	1.2	1	1	1.2	1	1	1.2	1	--	--	--
25	1	1.3	1	1	1.3	1	<b>1.52</b>	<b>0.11</b>	<b>2</b>	<b>1.52</b>	<b>0.11</b>	<b>2</b>	<b>0.092</b>	<b>0.035</b>	<b>0.064</b>
26	1	1.4	1	1	1.4	1	1	1.4	1	1	1.4	1	--	--	--
27	1	1.6	1	1	1.6	1	<b>2.00</b>	<b>-0.38</b>	<b>4</b>	<b>2.00</b>	<b>-0.38</b>	<b>4</b>	<b>0.123</b>	<b>0.100</b>	<b>0.093</b>
28	1	1.7	1	1	1.7	1	1	1.7	1	1	1.7	1	--	--	--
29	1	1.9	1	1	1.9	1	1	1.9	1	1	1.9	1	--	--	--
30	1	2.0	1	<b>2.40</b>	<b>-0.44</b>	<b>6</b>	<b>2.40</b>	<b>-0.44</b>	<b>6</b>	<b>2.40</b>	<b>-0.44</b>	<b>6</b>	<b>0.135</b>	<b>0.115</b>	<b>0.123</b>

Note, Items in bold indicate the items manipulated to include misfit

Table 5. 3PL Generating Model Parameters

Item	0% Misfitting			10% Misfitting			30% Misfitting			50% Misfitting			MISFIT	
	$a_i$	$b_i$	$c_i$	$a_i$	$b_i$	$c_i$	$a_i$	$b_i$	$c_i$	$a_i$	$b_i$	$c_i$	AM-2PL	AM-1PL
1	1	-2	0	1	-2	0	1	-2	0	1	-2	0	--	--
2	1	-1.9	0	1	-1.9	0	<b>1.45</b>	<b>-0.21</b>	<b>0.12</b>	<b>1.45</b>	<b>-0.21</b>	<b>0.12</b>	<b>0.019</b>	<b>0.022</b>
3	1	-1.7	0	1	-1.7	0	1	-1.7	0	<b>1.17</b>	<b>-0.52</b>	<b>0.13</b>	<b>0.018</b>	<b>0.019</b>
4	1	-1.6	0	1	-1.6	0	1	-1.6	0	1	-1.6	0	--	--
5	1	-1.4	0	1	-1.4	0	<b>1.53</b>	<b>1.75</b>	<b>0.25</b>	<b>1.53</b>	<b>1.75</b>	<b>0.25</b>	<b>0.070</b>	<b>0.081</b>
6	1	-1.3	0	1	-1.3	0	1	-1.3	0	<b>0.56</b>	<b>0.44</b>	<b>0.14</b>	<b>0.025</b>	<b>0.046</b>
7	1	-1.2	0	1	-1.2	0	1	-1.2	0	1	-1.2	0	--	--
8	1	-1.0	0	1	-1.0	0	1	-1.0	0	<b>1.48</b>	<b>0.14</b>	<b>0.13</b>	<b>0.025</b>	<b>0.022</b>
9	1	-0.9	0	1	-0.9	0	1	-0.9	0	1	-0.9	0	--	--

10	1	-0.8	0	<b>1.68</b>	<b>0.96</b>	<b>0.25</b>	<b>1.68</b>	<b>0.96</b>	<b>0.25</b>	<b>1.68</b>	<b>0.96</b>	<b>0.25</b>	<b>0.062</b>	<b>0.041</b>
11	1	-0.6	0	1	-0.6	0	1	-0.6	0	<b>1.32</b>	<b>0.26</b>	<b>0.14</b>	<b>0.027</b>	<b>0.022</b>
12	1	-0.5	0	1	-0.5	0	1	-0.5	0	1	-0.5	0	--	--
13	1	-0.3	0	1	-0.3	0	<b>2.40</b>	<b>-0.79</b>	<b>0.18</b>	<b>2.40</b>	<b>-0.79</b>	<b>0.18</b>	<b>0.022</b>	<b>0.046</b>
14	1	-0.2	0	1	-0.2	0	1	-0.2	0	1	-0.2	0	--	--
15	1	-0.1	0	1	-0.1	0	<b>2.01</b>	<b>0.04</b>	<b>0.20</b>	<b>2.01</b>	<b>0.04</b>	<b>0.20</b>	<b>0.038</b>	<b>0.036</b>
16	1	0.1	0	1	0.1	0	1	0.1	0	1	0.1	0	--	--
17	1	0.2	0	1	0.2	0	1	0.2	0	1	0.2	0	--	--
18	1	0.3	0	1	0.3	0	1	0.3	0	<b>1.53</b>	<b>1.75</b>	<b>0.25</b>	<b>0.070</b>	<b>0.051</b>
19	1	0.5	0	1	0.5	0	1	0.5	0	1	0.5	0	--	--
20	1	0.6	0	<b>1.70</b>	<b>0.40</b>	<b>0.16</b>	<b>1.70</b>	<b>0.40</b>	<b>0.16</b>	<b>1.70</b>	<b>0.40</b>	<b>0.16</b>	<b>0.034</b>	<b>0.025</b>
21	1	0.8	0	1	0.8	0	1	0.8	0	1	0.8	0	--	--
22	1	0.9	0	1	0.9	0	1	0.9	0	1	0.9	0	--	--
23	1	1.0	0	1	1.0	0	1	1.0	0	<b>0.60</b>	<b>-0.56</b>	<b>0.21</b>	<b>0.029</b>	<b>0.034</b>
24	1	1.2	0	1	1.2	0	1	1.2	0	1	1.2	0	--	--
25	1	1.3	0	1	1.3	0	<b>1.52</b>	<b>0.11</b>	<b>0.20</b>	<b>1.52</b>	<b>0.11</b>	<b>0.20</b>	<b>0.038</b>	<b>0.032</b>
26	1	1.4	0	1	1.4	0	1	1.4	0	1	1.4	0	--	--
27	1	1.6	0	1	1.6	0	<b>2.00</b>	<b>-0.38</b>	<b>0.17</b>	<b>2.00</b>	<b>-0.38</b>	<b>0.17</b>	<b>0.026</b>	<b>0.037</b>
28	1	1.7	0	1	1.7	0	1	1.7	0	1	1.7	0	--	--
29	1	1.9	0	1	1.9	0	1	1.9	0	1	1.9	0	--	--
30	1	2.0	0	<b>2.40</b>	<b>-0.44</b>	<b>0.40</b>	<b>2.40</b>	<b>-0.44</b>	<b>0.40</b>	<b>2.40</b>	<b>-0.44</b>	<b>0.40</b>	<b>0.061</b>	<b>0.066</b>

*Note, Items in bold indicate the items manipulated to include misfit*

Table 6. 2PL Generating Model Parameters

Item	0% Misfitting		10% Misfitting		30% Misfitting		50% Misfitting		MISFIT AM-1PL
	$a_i$	$b_i$	$a_i$	$b_i$	$a_i$	$b_i$	$a_i$	$b_i$	
1	1	-2	1	-2	1	-2	1	-2	--
2	1	-1.9	1	-1.9	<b>1.45</b>	<b>-0.21</b>	<b>1.45</b>	<b>-0.21</b>	<b>0.019</b>
3	1	-1.7	1	-1.7	1	-1.7	<b>1.17</b>	<b>-0.52</b>	<b>0.008</b>
4	1	-1.6	1	-1.6	1	-1.6	1	-1.6	--
5	1	-1.4	1	-1.4	<b>1.53</b>	<b>1.75</b>	<b>1.53</b>	<b>1.75</b>	<b>0.022</b>
6	1	-1.3	1	-1.3	1	-1.3	<b>0.56</b>	<b>0.44</b>	<b>0.026</b>
7	1	-1.2	1	-1.2	1	-1.2	1	-1.2	--
8	1	-1.0	1	-1.0	1	-1	<b>1.48</b>	<b>0.14</b>	<b>0.019</b>
9	1	-0.9	1	-0.9	1	-0.9	1	-0.9	--
10	1	-0.8	<b>1.68</b>	<b>0.96</b>	<b>1.68</b>	<b>0.96</b>	<b>1.68</b>	<b>0.96</b>	<b>0.027</b>
11	1	-0.6	1	-0.6	1	-0.6	<b>1.32</b>	<b>0.26</b>	<b>0.014</b>
12	1	-0.5	1	-0.5	1	-0.5	1	-0.5	--
13	1	-0.3	1	-0.3	<b>2.40</b>	<b>-0.79</b>	<b>2.40</b>	<b>-0.79</b>	<b>0.044</b>
14	1	-0.2	1	-0.2	1	-0.2	1	-0.2	--
15	1	-0.1	1	-0.1	<b>2.01</b>	<b>0.04</b>	<b>2.01</b>	<b>0.04</b>	<b>0.036</b>
16	1	0.1	1	0.1	1	0.1	1	0.1	--
17	1	0.2	1	0.2	1	0.2	1	0.2	--
18	1	0.3	1	0.3	1	0.3	<b>1.53</b>	<b>1.75</b>	<b>0.022</b>
19	1	0.5	1	0.5	1	0.5	1	0.5	--
20	1	0.6	<b>1.70</b>	<b>0.40</b>	<b>1.70</b>	<b>0.40</b>	<b>1.70</b>	<b>0.40</b>	<b>0.027</b>
21	1	0.8	1	0.8	1	0.8	1	0.8	--
22	1	0.9	1	0.9	1	0.9	1	0.9	--
23	1	1.0	1	1.0	1	1.0	<b>0.60</b>	<b>-0.56</b>	<b>0.024</b>
24	1	1.2	1	1.2	1	1.2	1	1.2	--
25	1	1.3	1	1.3	<b>1.52</b>	<b>0.11</b>	<b>1.52</b>	<b>0.11</b>	<b>0.021</b>
26	1	1.4	1	1.4	1	1.4	1	1.4	--
27	1	1.6	1	1.6	<b>2.00</b>	<b>-0.38</b>	<b>2.00</b>	<b>-0.38</b>	<b>0.036</b>
28	1	1.7	1	1.7	1	1.7	1	1.7	--
29	1	1.9	1	1.9	1	1.9	1	1.9	--
30	1	2.0	<b>2.40</b>	<b>-0.44</b>	<b>2.40</b>	<b>-0.44</b>	<b>2.40</b>	<b>-0.44</b>	<b>0.044</b>

**Sample size.** Several sources have noted that, with small sample sizes, noninformative priors have a more profound effect on the estimation of the posterior distributions, which in turn can affect the PPC procedure (Berkhof, et al. 2006; Lambert, 2005; O'Hagan, 1994). Typical studies of PPC for IRT have used very large sample sizes. For instance, Sinharay and Johnson (2003); Sinharay (2006); and Sinharay, Johnson, and Stern (2006) used 2500 individuals. At such large sample sizes, it is anticipated that the role of the prior will be negligible. However, there is a lack of information in the literature regarding the performance of PPC using smaller sample sizes and to what extent the prior specification will affect the method. Sinharay (2006) performed a simulation with 500, 1000, 2000, and 4000 individuals. Some other sample sizes were investigated by Toribio and Albert (2011), who used 1000 individuals. However, Levy et al. (2009) investigated PPC for multidimensional IRT using 250, 750, and 2500 individuals. A parameter recovery study of the unidimensional 3PL by Sheng (2010) used 100, 300, 500, and 1000 individuals. In practical terms, sample sizes under 500 would rarely be used when fitting a 3PL in an operational setting. Therefore, sample sizes of 500, 1000, and 2500 individuals will be used in the current study.

**Discrepancy statistics.** In the PPC process, if the IRT model is a good fit to the data, then future item response data simulated from the model should look very much like observed data. Conversely, if the model is a poor fit, then future simulated data will look different from observed data (Lynch, 2007). Determining how similar the simulated and observed data are requires choice of an indicator, often termed discrepancy statistic to

highlight the focus on measuring discrepancies, or differences, between a model and the data (Meng, 1994).

The literature does not show a consensus as to the best choice of discrepancy statistic. For example, Sinharay, Johnson, and Stern (2003) examined a comparison of observed-score and predicted-score distributions to assess fit of the model to the data. Sinharay (2006) used  $S-X^2$  and  $S-G^2$ , showing their type I error rates and false alarm rates did not exceed the nominal level. However, the tests were also shown to be conservative, failing to detect some misfitting items. Toribio and Albert (2011) examined *OUTFIT*, *INFIT*, Bock's Pearson-type  $\chi^2$  index (1972),  $Q_1$ ,  $G^2$ ,  $S-X^2$  and  $S-G^2$ , with some conflicting results from those reported by Sinharay (2006) (possibly caused by the alternative computation method used by Toribio and Albert for  $S-X^2$  and  $S-G^2$ ).

Since the focus of this work is not on the performance of individual discrepancy measures, only a selected few which have been studied in previous simulations will be used. For example, in the Frequentist framework, Sinharay (2006) concluded that  $S-X^2$  is the best choice and provides acceptable, if not slightly conservative, false positive error rates. This study will examine the percent correct (p-value),  $S-X^2$ , and *INFIT*, in part because these have been examined in previous studies, but also because a value as simple as percent correct could greatly increase the speed with which a PPC procedure could be performed.

**Choice of priors.** Multiple studies (Sinharay et al., 2006; Sinharay, 2005; Sinharay & Johnson, 2003) have used noninformative priors for dichotomous models such as  $\log(a_i) \sim N(0, 10)$  and  $b_i \sim N(0, 10)$ . Sinharay (2006) used more informative

reference priors of  $\log(a_i) \sim N(0, 1)$ ,  $b_i \sim N(0, 1)$ , and  $\text{logit}(c_i) \sim \text{NIID}(\text{logit}(0.2))$ .

Sheng (2010) varied the informativeness of the parameter priors  $a_i$  and  $b_i$  in several ways:

(a) noninformative uniform prior; (b) noninformative Normal prior with a large variance of  $10^{10}$ ; (c) more informative prior, with a variance of 4; and (d) informative prior with a variance of 1. Doing this, Sheng found that relatively informative priors, when

accurately specified, should be adopted for the discrimination and difficulty parameters.

Sheng also considered three priors for guessing: (a) noninformative Beta prior ( $\text{Beta}(1, 1)$ ); (b) informative Beta prior with mean 0.22 and standard deviation of 0.131 ( $\text{Beta}(2, 7)$ ); and (c) very precise Beta prior with mean 0.22 with standard deviation of 0.007 ( $\text{Beta}(5, 17)$ ).

Sinharay and Johnson (2003) comment that strongly informative priors may seriously affect the results of posterior predictive checks, and that the replicated data sets obtained under strong, and incorrect prior distributions may be systematically far from observed data. The result would be a very large or very small PPP value, which may lead the researcher to conclude incorrectly that model-data misfit exists. However, a strong and accurate prior distribution can help the researcher assess the fit of the likelihood more effectively (Gelman et al., 1996; Sinharay & Johnson, 2003). Sheng (2010) followed a different procedure and investigated each specification in isolation. For example, when the estimation of the discrimination parameter for the 3PNO was of interest, the prior variance was manipulated for the discrimination parameter only, keeping the difficulty and guessing parameters noninformative. However, it appears more common to specify all the parameters as either noninformative or informative (e.g., Levy et al., 2009;

Sinharay, 2005; Sinharay, 2006; Sinharay et al., 2006; Sinharay & Johnson, 2003). These studies will guide my choice of the prior specifications. All parameters will be specified in a similar fashion such that if one parameter is noninformative, all parameters are noninformative. The prior specifications are found in Table 7.

Table 7. Prior Specification

Prior Specification	Discrimination	Difficulty	Guessing
Noninformative	$a_i \sim N_{(0,\infty)}(0, 100)$	$b_i \sim N(0,100)$	$c_i \sim Beta(1, 1)$
Informative-Accurate	$a_i \sim N_{(0,\infty)}(1, 1)$	$b_i \sim N(\mu_b, 1)$ , where $\mu_b$ reflects the percent correct (p-value) for the item placed on a similar metric as the IRT scale $(-\ln(\frac{percent\ correct_i}{1-percent\ correct_i}))$ .	$c_i \sim Beta(2, 7)$
Informative-Inaccurate	$a_i \sim N_{(0,\infty)}(3, 1)$	$b_i \sim N(\mu_b, 1)$ , where $\mu_b$ reflects the percent correct (p-value) for the item placed on a similar metric as the IRT scale, but inaccurate $(\ln(\frac{percent\ correct_i}{1-percent\ correct_i}))$	$c_i \sim Beta(5, 4)$
Noninformative-Inaccurate	$a_i \sim N(0, 100)$	$b_i \sim N(\mu_b, 100)$ , where $\mu_b$ reflects the percent correct (p-value) for the item placed on a similar metric as the IRT scale, but inaccurate $(\ln(\frac{percent\ correct_i}{1-percent\ correct_i}))$	$c_i \sim Beta(10, 7)$

In summary, Table 8 contains the conditions proposed to answer the research questions.

Table 8. Simulation Conditions

<i>Condition</i>	<i>Values</i>
Percent of misfitting items	0%, 10%, 30%, 50%
Sample size	500, 1000, 2500
Prior Specification	Noninformative, Informative-Accurate, Informative-Inaccurate, Noninformative-Inaccurate
Degree of misfit	Small (MISFIT < 0.020), Medium to Large (MISFIT ≥ 0.020)
Type of misfit introduced	GM-LPE with AM-3PL, 2PL, 1PL; GM-3PL with AM-2PL, 1PL; GM-2PL with AM-1PL

### **Estimation**

Each data set will contain responses of  $N$  examinees to  $I$  dichotomous items. The generating parameter values (for items and examinees) are the same for all 100 data sets within a GM and are shown in Tables 4 through 6. For each of the 100 data sets generated, the AM is fit to the data using an MCMC algorithm. The MCMC algorithm will be implemented in SAS version 9.4 (SAS Institute, Inc., 2013), per Ames and Samonte (in press). A review of the MCMC procedure to be implemented in SAS is provided below.

As described in Chapter 2, MCMC is a method of sampling which uses a Markov chain and the Monte Carlo principle (Jackman, 2009) to sample from a posterior distribution which is not a common type. Once constructed, the Markov chain represents the posterior distribution and each point in the chain represents a sample of the posterior. For a more technical and detailed description of MCMC for IRT, see Patz and Junker (1999) and Kim and Bolt (2007).

The Metropolis-Hastings (MH; Hastings, 1970; Metropolis et. al., 1953; Metropolis & Ulam, 1949) algorithm is one method for creating the Markov chain sequence. I will use the MH algorithm because SAS PROC MCMC implements the random walk MH algorithm. The chain starts with a beginning value of the parameter,  $b^{start}$ , which can be a completely random value or a value specified by the practitioner to represent a belief about the parameter. The algorithm transitions to the next value of the parameter (i.e., the next element of the chain),  $b^{next}$ , using the following general steps for this algorithm:

1. Begin with  $b^{next}$ , a candidate value for the Markov chain, drawn from a *proposal distribution*. A popular proposal distribution is the “random walk” proposal. That is, a value is randomly sampled from somewhere near the current point ( $b^{next} = b^{start} + \text{random error}$ ). The proposal distribution can change over iterations, a process referred to as *tuning*.
2. Compute an acceptance ratio,  $r$ , which represents the plausibility of the candidate value,  $b^{next}$ . The acceptance ratio ranges from 0 to 1 and represents how likely the candidate value is to have come from the posterior. Low acceptance ratios indicate the candidate value is not likely to have come from the posterior. High acceptance ratios indicate the candidate value is quite likely to have come from the posterior.
3. If  $r > 1$ , move to the candidate value,  $b^{next}$ . Otherwise, move to the candidate value with probability,  $r$ , and remain at the existing value,  $b^{start}$ , with probability  $1-r$ .

4. Repeat a specified number of times. The exact number of repetitions is specified by the practitioner. With each repetition of the algorithm, or iteration, the starting value is either the previous iteration's accepted candidate or the previous iteration's starting value. Once enough iterations have been completed, the sequence represents values of the posterior distribution and the information can be summarized.

Depending on the starting value, the first several elements of the chain are typically not very good representations of the parameter and are thrown away. These throw-away elements are called the *burn-in*. As the algorithm continues through the parameter space, the algorithm will narrow in on one general location, the central mass of the distribution. Information about each instance of the algorithm, which provides information about the posterior, is saved.

Extending the MH algorithm to a higher-dimensional parameter space is straightforward (Chib and Greenberg (1995) provide a tutorial on the algorithm for multiple dimensions). Let the  $k$ -dimensional parameter vector be represented by  $\boldsymbol{\eta} = (\eta^1, \dots, \eta^k)$ . An initial start value for each  $\eta^k$  is chosen and a multivariate version of the random walk proposal distribution, such as a multivariate normal distribution, is used to select a  $k$ -dimensional new parameter. Other steps remain the same as those previously described.

In keeping with the procedure used by Sinharay (2006), an initial 10,000 iterations will be run after a burn-in of 2,000 (total iterations=12,000). However, these

values may be adjusted depending upon the convergence of the posterior. A standard convergence statistic reported by SAS PROC MCMC is the Geweke diagnostic (Geweke, 1992), which assesses whether the posterior estimates have converged by comparing means from the early and latter part of the Markov chain. This is accomplished via a two-sided test based on a z-score statistic. If the 10,000 iterations are not enough to achieve convergence, the values will be increased until convergence is achieved. For 1000 draws in the final posterior sample, a replicated data set will be simulated and values of the realized and predictive discrepancy measures will be computed. PPP values will be computed for each of the simulation conditions.

### **Outcomes**

The two outcomes of interest will be the Type I error rate and Power rate, for each discrepancy measure. If an item is simulated to contain misfit and the PPP value is less than .05 or greater than .95, then the item is correctly identified as misfitting. However, if an item is not simulated to contain misfit, and the PPP value is less than .05 or greater than .95, then the item is incorrectly identified as misfitting. The proportion of items incorrectly identified as misfitting is the Type I error rate. If an item is not simulated to contain misfit, and the PPP value is between .05 and .95 (inclusive), then the item is correctly identified as fitting well. The proportion of items correctly identified as fitting well is the Power rate.

### **Analysis**

A fully crossed, four-way ANOVA (prior specification as one factor, sample size as the second factor, percent of misfitting items as the third factor, and degree of misfit as

the final factor) will be used to determine whether differences in Type I error and Power values exist given the factors of interest for each discrepancy measure. The four-way ANOVA will be performed for each of the three discrepancy measures ( $S-X^2$ , *INFIT*, percent correct).

## **CHAPTER IV**

### **RESULTS**

This chapter presents results from the simulation study. The simulation was designed to address the four research questions posited at the end of Chapter One. To summarize the problem and literature described in Chapter Two: the studies in PPC lack consistency with respect to how priors have been specified. The PPC method might be sensitive to choices in specification of the prior distributions because of the flatness of the likelihood in the presence of model-data misfit, specifically the likelihood will show decreased curvature as model-data misfit increases. With a decreased curvature, the prior will have more of an effect on posterior distributions and PPD. However, it is common for researchers to use noninformative priors in this approach, relying on the belief that the noninformative prior will have little effect on the procedure (Sinharay, 2006).

The methods detailed in Chapter Three provided guidance in assessing the PPC approach's sensitivity to prior specification, as well as under which conditions researchers must be most attentive to the choice of priors for PPC. Conditions tested in the study design included sample size (500, 1000, 2500), percent of misfitting items (0%, 10%, 30%, 50%), degree of misfit (small, medium-large), and the type of misfit introduced when the GM and AM differed.

The results are organized as follows: First, I present a discussion of the convergence diagnostics and the time to converge. Second, I address the results pertaining to each of the four research questions, specifically:

1. To what extent does prior specification influence the results of the PPC method for the model-data fit of unidimensional, dichotomous IRT models?
2. How does sample size affect the influence of prior specification on the results of the PPC method for the model-data fit of unidimensional, dichotomous IRT models?
3. How does the type of misfit affect the influence of prior specification on the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?
4. How does the interaction of sample size and type of misfit affect the influence of prior specification on the results of the PPC method for model-data fit of unidimensional, dichotomous IRT models?

I used false positive error rates and hit rates as measures of how the PPC method performs.

### **Convergence**

The Geweke diagnostic (Geweke, 1992) is a standard convergence statistic reported by SAS PROC MCMC, and it assesses whether the posterior estimates have converged by comparing means from the early and latter part of the Markov chain. This is accomplished via a two-sided test based on a z-score statistic - specifically, two, non-overlapping sequences of the Markov chain (i.e., the parameter posterior distribution) for

any parameter, such as  $\theta$ , are considered ( $\theta_1^t: t = 1, \dots, iter_1; \theta_2^t: t = iter_a, \dots, iter$  where  $1 < iter_1 < iter_a < iter$  and  $iter =$  iteration of the Markov chain). Let

$$\bar{\theta}_1 = \frac{1}{iter_1} \sum_{t=1}^{iter_1} \theta_1^t \text{ and } \bar{\theta}_2 = \frac{1}{iter_2} \sum_{t=iter_a}^{iter} \theta_2^t, \quad (20)$$

where  $iter_2 = iter - iter_a + 1$ . If the ratios  $iter_1/iter$  and  $iter_2/iter$  are fixed,  $\frac{iter_1+iter_2}{n} < 1$ , and the chain is stationary, then the following statistic converges to a standard normal distribution as the number of iterations approaches infinity:

$$Z_{iter} = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\frac{\hat{s}_1(0) + \hat{s}_2(0)}{iter_1 + iter_2}}}. \quad (21)$$

In Equation 21,  $\hat{s}_1(0)$  and  $\hat{s}_2(0)$  denote consistent spectral density estimates at zero frequency. Large absolute values of the Geweke Z statistic indicate rejection of the null hypothesis of convergence.

In addition to the Geweke diagnostic, I also used visual analysis of trace plots to assess convergence. Trace plots of posterior samples on the vertical axis against the iteration index on the horizontal axis can be very useful in assessing convergence. The trace plot is used in conjunction with Geweke statistics to determine if the chain has not yet converged to its stationary distribution, which would indicate whether the chain needs a longer burn-in period and whether the chain is mixing well. The aspects of convergence that are most recognizable from a trace plot are a relatively constant mean and variance.

**Noninformative, Informative-Accurate, and Informative-Inaccurate Priors.**

In keeping with the procedure used by Sinharay (2006), I ran an initial 10,000 iterations after a burn-in of 2,000 (total iterations=12,000). However, I subsequently modified these values, based on sample runs of one replicate per condition, for the condition when the percent of misfitting items was 50%, there were 2500 people, the prior specification was noninformative, GM=LPE, and AM=3PL. After increasing the number of iterations incrementally, by 1,000 iterations at a time, a total of 12,000 iterations were used after a burn-in of 5,000 iterations (total iterations=17,000) for the remainder of the simulation.

**Noninformative-Inaccurate Priors.** For MCMC algorithms using a noninformative-inaccurate prior distribution, the number of tuning loops reached 26 (the maximum value of the MH algorithm in SAS PROC MCMC). Tuning is the process of finding a good proposal distribution for each block of parameters. The tuning phase consists of a number of loops and each loop lasts for a number of iterations. By default, SAS PROC MCMC uses 500 iterations at the end of each loop. At the end of every loop, the acceptance probability for each parameter block is examined. If the probability falls within the acceptance tolerance range, the current proposal distribution is kept. Otherwise, the proposal distribution parameters are modified before the next tuning loop.

For the simulations using a noninformative-inaccurate prior distribution, some parameters had acceptance probabilities either outside of the target range of 0.159 to 0.309 or below the target probability of 0.6. This indicates that the proposal distributions were not fully tuned, resulting in potentially bad mixing of the Markov chain.

Initially, even after performing 5,000 burn-in iterations and 12,000 iterations to generate the Markov chain samples, Geweke diagnostics could not be computed because the variances of the two segments used in computation of the Geweke diagnostic were both 0. This was caused by the Markov chain being a constant vector. I then doubled both the number of burn-ins and replications (to 10,000 burn-ins and 24,000 Markov chain samples, respectively), but the resulting proposal distributions were still not fully tuned. Model-checking would not occur if this were the case. Hence, the remaining results are discussed only for the noninformative, informative-accurate, and informative-inaccurate prior distributions, and not for the noninformative-inaccurate priors.

### **Time to Converge**

Previous research (Sheng, 2010) has found that the informativeness of priors affects not only the convergence of Markov chains, but also the time needed for the MCMC algorithm to converge. Time to converge was therefore monitored because it can be a factor in adoption of MCMC methods. Figure 7 provides a comparison of the average time to converge, across prior informativeness.

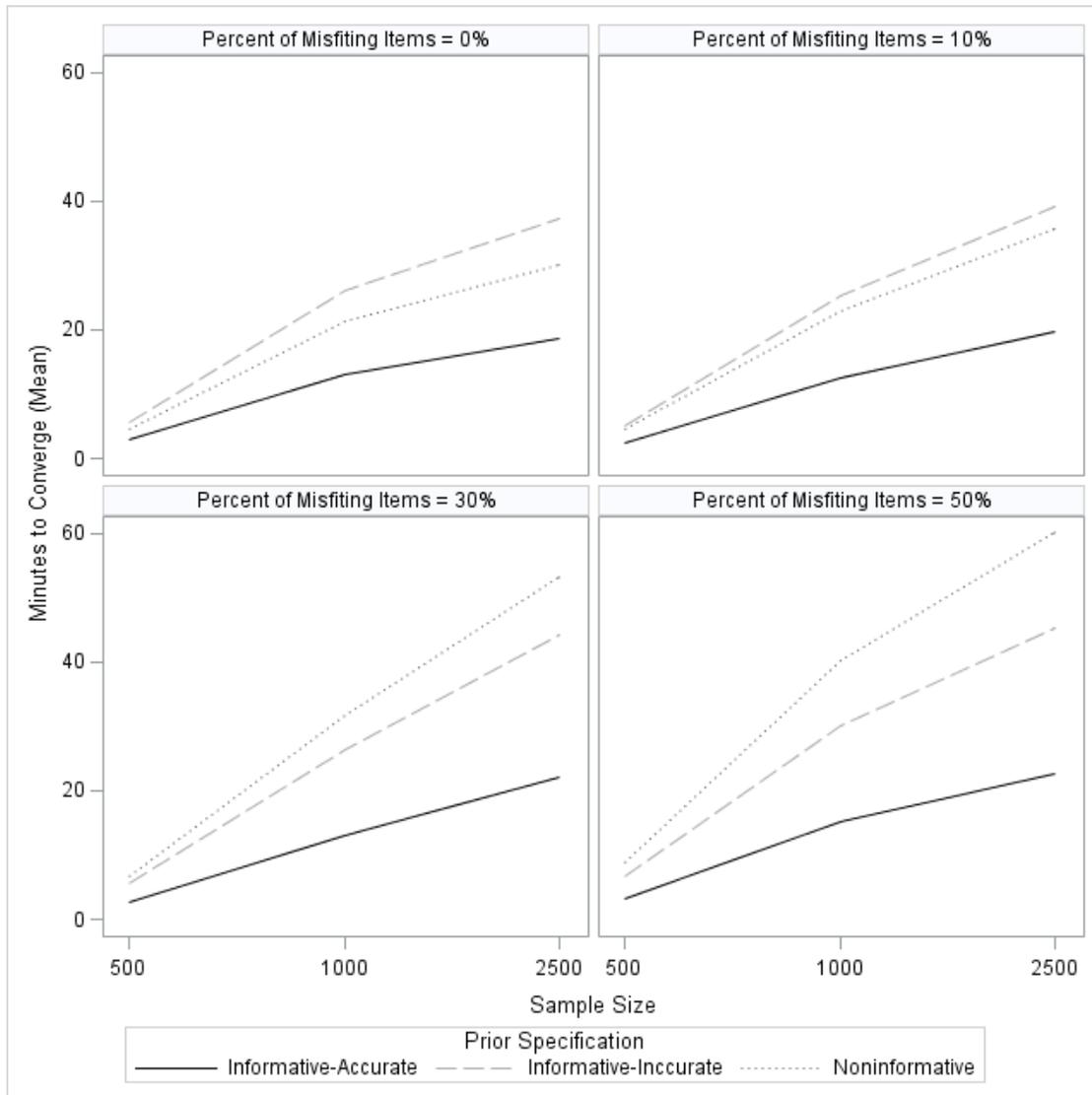


Figure 7. Average Time to Converge, in minutes, by Type of Prior

As illustrated in Figure 7, in all instances, the use of informative-accurate priors resulted in faster times to converge than when noninformative or informative-inaccurate priors were used. When the percent of misfitting items was low (0% or 10% misfitting), noninformative priors tended to result in slightly faster convergence times than those of informative-inaccurate priors. However, when the percent of misfitting items was high

(30% or 50% misfitting), noninformative priors tended to result in slower convergence times than those of informative-inaccurate priors. Convergence times, compared across priors, were closest together with smaller sample sizes, and distributed farther apart for larger sample sizes. In all cases, time to converge increased as sample size increased because the MCMC algorithm estimates a person ability parameter posterior. Thus, as sample size increases, so too does the number of parameter posterior distributions.

### **Analysis**

Three discrepancy statistics ( $S-X^2$ , *INFIT*, and percent correct) were evaluated separately for false positive error rates (similar to Type I error, a false positive error occurs when an item is flagged as misfitting when it does not contain simulated misfit) and hit rates (similar to power, a hit occurs when an item is correctly flagged when it was generated to be misfitting).

An example of both of these terms – hit rate and false positive rate - is shown in Table 5 and the condition with 50% misfit: Item 1 was not generated to be a misfitting item, so if the PPC procedure indicates the item is misfitting, this is considered a false positive; conversely, Item 2 is generated to have misfit, so if the PPC procedure indicates the item is misfitting, this is considered a hit.

Results for each research question are discussed separately, and are further divided by item being evaluated (false positive error rates, hit rates) and by discrepancy statistic. Fully crossed factorial analysis of variance (ANOVA) tests were performed to investigate mean differences among hit rates and false positive error rates. The false positives and hits were averaged across 100 replications to create mean hit rates. Because

of the large data set size, the significance level for the ANOVA tests was set to .001. For the multiple comparison procedure, Tukey’s Honest Significant Difference (HSD; Tukey, 1949) was used and a significance level of .001 was again set to indicate significant mean differences.

**Research Question 1: To What Extent Does Prior Specification Influence the Results of the PPC Method for the Model-data Fit of Unidimensional, Dichotomous IRT models?**

This section provides results of the first research question.

**Hit Rate**

Table 9 provides a summary of the hit rate results for research question 1. Also included in Table 9 are partial eta-squared effect sizes ( $\eta_p^2$ ). These effect sizes can be interpreted as follows. Consider the  $\eta_p^2$  for *INFIT* in Table 9 of .062. This would indicate that 6.2% of the variation in the hit rates outcome is explained by prior specification. Suggested heuristics for interpreting partial eta-squared are provided by Cohen (1988) as small = 0.01; medium = 0.06; and large = 0.14. Thus, the effect size for *INFIT* is considered medium whereas the effect size for *S-X<sup>2</sup>* is considered small and considered small to medium for percent correct.

Table 9. Mean Hit Rates by Prior Specification

	Informative-Accurate	Informative-Inaccurate	Noninformative	Significance	$\eta_p^2$
<i>S-X<sup>2</sup></i>	.35	.37	.35		.002
<i>INFIT</i>	.63	.73	.44	***	.062
Percent correct	.03	.07	.01	***	.016

Note, \*\*\* indicates statistical significance of the one-way ANOVA at the .001 level.

**S-X<sup>2</sup>.** As indicated in Table 9, a one-way ANOVA was performed on mean hit rates for misfitting items. Using  $S-X^2$  as the discrepancy statistic, the ANOVA was not significant ( $F(2, 2913) = 0.26, p = .7737$ ) for the main effect of prior informativeness. This indicates that the mean hit rates were not significantly different either for noninformative priors ( $M_{hit} = 0.35$ ), informative-accurate priors ( $M_{hit} = 0.35$ ) or informative-inaccurate priors ( $M_{hit} = 0.37$ ).

**INFIT.** Using *INFIT* as the discrepancy statistic, the ANOVA was significant ( $F(2, 2913) = 96.35, p < .0001$ ) for the main effect of prior informativeness, indicating that the mean hit rates were significantly different for noninformative priors ( $M_{hit} = 0.44$ ), informative-accurate priors ( $M_{hit} = 0.63$ ) and informative-inaccurate priors ( $M_{hit} = 0.73$ ); in more detail, the hit rate for noninformative priors was found to be significantly lower than the hit rate for informative-accurate priors ( $M_{hit\_diff} = 0.19$ , where *hit\_diff* designates the difference in hit rates) and significantly lower than the hit rate for informative-inaccurate priors ( $M_{hit\_diff} = 0.29$ ). Further, the hit rate for informative-accurate priors was found to be significantly lower than the hit rate for informative-inaccurate priors ( $M_{hit\_diff} = 0.10$ ).

**Percent correct.** Using percent correct as a discrepancy statistic, the ANOVA was significant ( $F(2, 2913) = 23.44, p < .0001$ ). This indicates that the mean hit rates were significantly different for noninformative priors ( $M_{hit} = 0.01$ ), informative-accurate priors ( $M_{hit} = 0.07$ ) and informative-inaccurate priors ( $M_{hit} = 0.03$ ); more precisely, the hit rate for noninformative priors was found to be significantly lower than the hit rate for informative-accurate priors ( $M_{hit\_diff} = 0.06$ ) and significantly lower than the hit rate for

informative-inaccurate priors ( $M_{hit\_diff} = 0.02$ ). The hit rate for informative-accurate priors was also significantly lower than the hit rate for informative-inaccurate priors ( $M_{hit\_diff} = 0.04$ ).

### False Positive Error

Table 10 provides a summary of the false positive rate results for research question 1. For  $S-X^2$  and *INFIT*, false positive rates are elevated above the traditional .05 level. Because the PPC procedure flags an item as misfitting is the PPP value is below .05 or above .95, the type I error for this case is .10 rather than .05. Thus, the false positive rates in this chapter should be compared to .10. Again, the effect size partial eta-squared is included in Table 10.

Table 10. Mean False Positive Rates by Prior Specification

	Informative- Accurate	Informative- Inaccurate	Noninformative	Significance	$\eta_p^2$
$S-X^2$	.19	.20	.19		.000
<i>INFIT</i>	.23	.34	.19	***	.028
Percent correct	.02	.04	.01	***	.006

Note, \*\*\* indicates statistical significance of the one-way ANOVA at the .001 level.

$S-X^2$ . Using  $S-X^2$  as the discrepancy statistic, the one-way ANOVA on false positive error rates was not significant ( $F(2, 10041) = 0.02, p = .9766$ ) for the main effect of prior informativeness, indicating that the mean false positive error rates were not significantly different for noninformative priors ( $M_{false+} = 0.19$ ), informative-accurate priors ( $M_{false+} = 0.19$ ) or informative-inaccurate priors ( $M_{false+} = 0.20$ ). However, all were greater than .10, the type I error for the PPC procedure.

**INFIT.** Using *INFIT* as the discrepancy statistic, the ANOVA on false positive error rates was significant ( $F(2, 10041) = 144.68, p < .0001$ ). This indicates that the mean false positive error rates were significantly different for noninformative priors ( $M_{false+} = 0.19$ ), informative-accurate priors ( $M_{false+} = 0.23$ ) and informative-inaccurate priors ( $M_{false+} = 0.34$ ) - specifically, the false positive error rate for noninformative priors was found to be significantly lower than the false positive error rate for informative-accurate priors ( $M_{false+_{diff}} = 0.04$ ) and significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.15$ ). In addition, the false positive error rate for informative-accurate priors was found to be significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.11$ ). Similar to the  $S-X^2$  ANOVA results, all false positive error rates were elevated above the desirable level of 0.10.

**Percent correct.** Using percent correct as a discrepancy statistic, the ANOVA was significant ( $F(2, 10041) = 31.71, p < .0001$ ). This indicates that the mean false positive error rates were significantly different for noninformative priors ( $M_{false+} = 0.01$ ), informative-accurate priors ( $M_{false+} = 0.02$ ) and informative-inaccurate priors ( $M_{false+} = 0.04$ ) - specifically, the false positive error rate for noninformative priors was found to be significantly lower than the false positive error rate for informative-accurate priors ( $M_{false+_{diff}} = 0.01$ ) and significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.03$ ). The false positive error rate for informative-accurate priors was also found to be significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.02$ ).

## Summary

Significantly different mean hit rates for each type of prior used were found when using *INFIT* and percent correct as discrepancy statistics. *INFIT* had the highest overall hit rates, peaking when using informative-inaccurate priors. Using percent correct had the lowest overall hit rates, and was at its lowest with noninformative priors. Figure 8 illustrates mean hit rates for each discrepancy statistic, across the different types of priors.

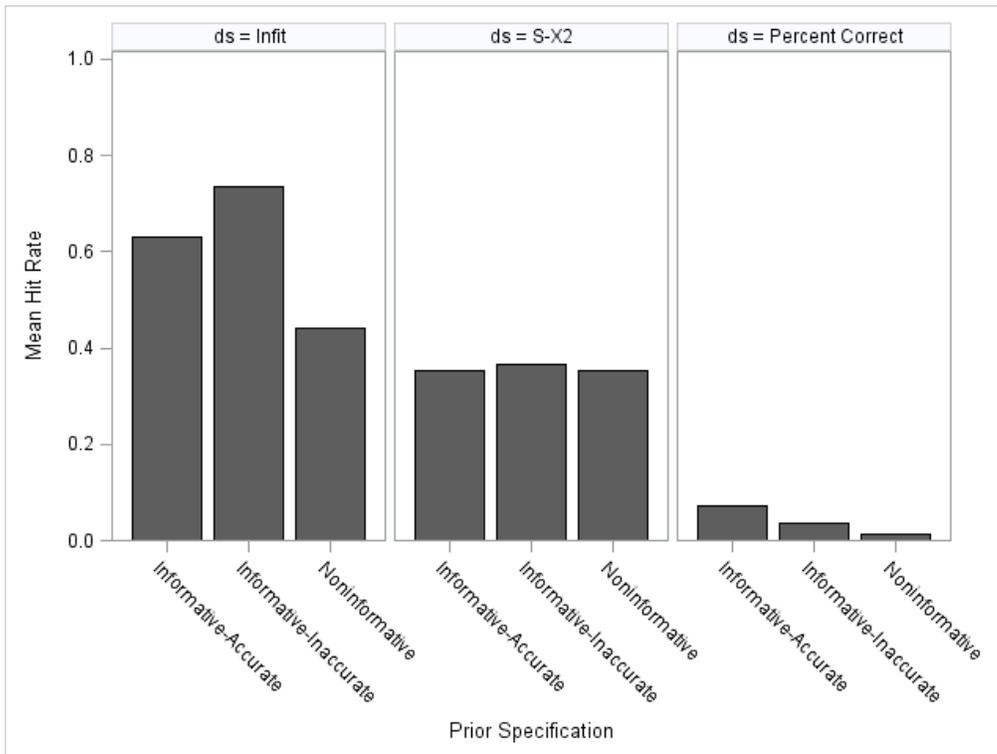


Figure 8. Mean Hit Rates, by Type of Prior

Significantly different mean false error positive rates, for each type of prior used, were found when using *INFIT* and percent correct as discrepancy statistics. *INFIT* had the

highest overall false positive error rates, peaking when using informative-inaccurate priors. Figure 9 illustrates mean false positive error rates for each discrepancy statistic, across the different types of priors. Based on these results, informative-inaccurate priors tend to categorize most items as misfitting, regardless of whether the item was generated to contain misfit or not.

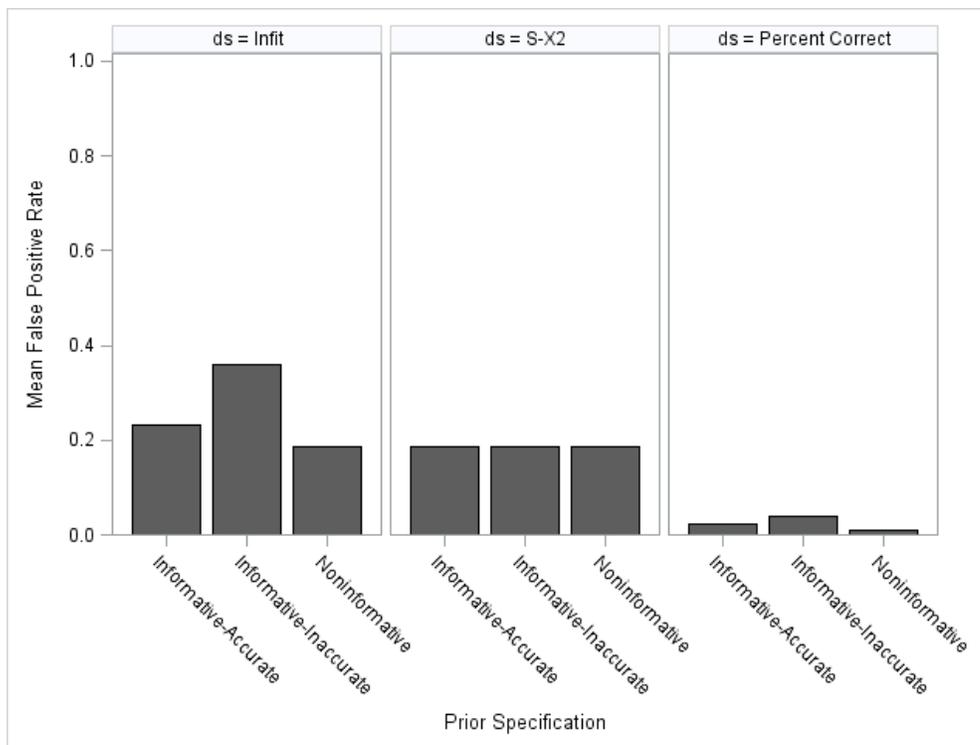


Figure 9. Mean False Positive Rates, by Type of Prior

**Research Question 2: How Does Sample Size Affect the Influence of Prior Specification on the Results of the PPC Method for Model-data Fit of Unidimensional, Dichotomous IRT Models?**

This section provides results of the second research question.

## Hit Rate

Table 11 provides a summary of the mean hit rates for Research Question 2 and Figure 10 provides a summary of these results for all three discrepancy statistics.

Table 11. Mean Hit Rates by Prior Specification and Sample size

Prior	Sample size	$S-X^2$	<i>INFIT</i>	Percent Correct***
Informative-Accurate	500	.20	.54	.01
	1000	.35	.60	.05
	2500	.52	.75	.16
Informative-Inaccurate	500	.23	.65	.02
	1000	.35	.72	.04
	2500	.52	.83	.05
Noninformative	500	.18	.40	.01
	1000	.35	.42	.01
	2500	.54	.50	.02

Note, \*\*\* indicates statistical significance of the two-way ANOVA at the .001 level.

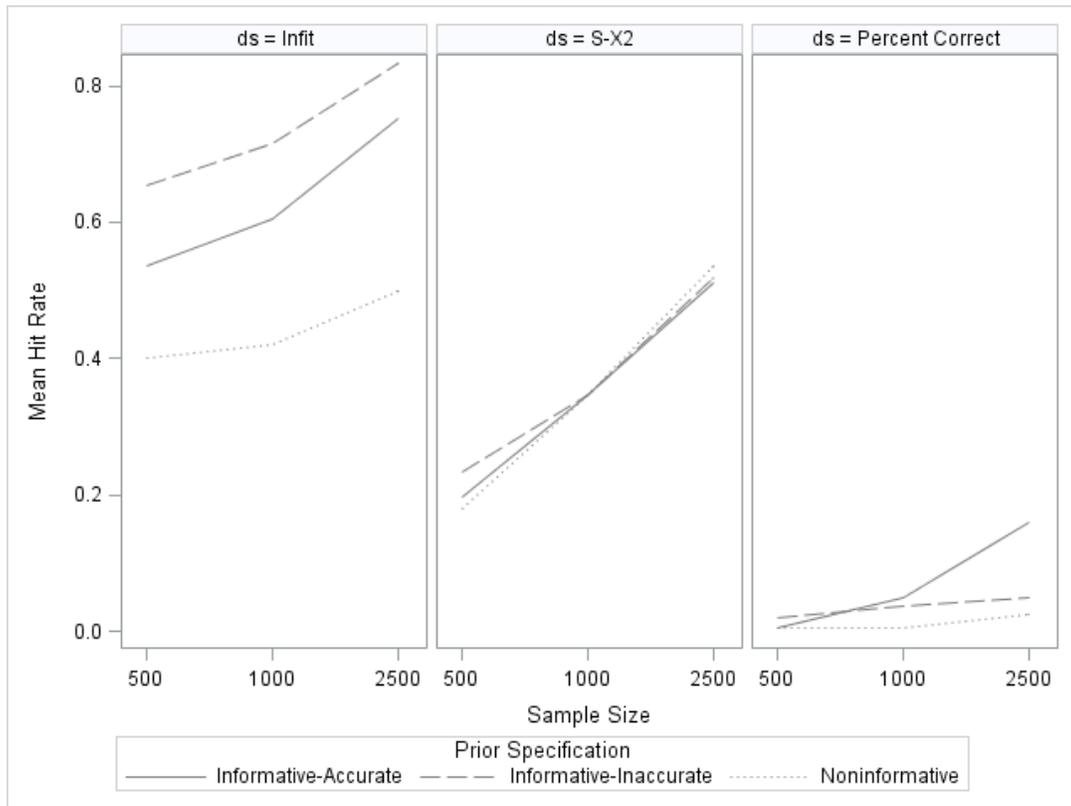


Figure 10. Mean Hit Rates, by Type of Prior, Sample size, and Discrepancy Statistic

Partial eta-squared effect sizes for research questions two through four are found in Table 12, for both main effects and interactions.

Table 12. Effect Sizes for Research Questions Two Through Four

Research Question	Effect	Hit Rates			False Positive Rates		
		$\eta_p^2$			$\eta_p^2$		
		S- $X^2$	INFIT	Percent Correct	S- $X^2$	INFIT	Percent Correct
Two	Prior	.000	.063	.017	.000	.028	.007
	Sample size	.074	.021	.022	.002	.026	.023
	Prior*Sample size	.001	.002	.019	.000	.000	.006
Three	Prior	.000	.073	.017			
	Size	.029	.004	.001			
	Location	.044	.112	.031			
	Prior*Size	.000	.003	.000			
	Prior*Location	.000	.050	.015			
	Prior*Size*Location	.002	.011	.002			
Four	Prior	.000	.076	.018			
	Size	.033	.004	.001			
	Location	.049	.117	.033			
	Sample size	.082	.026	.024			
	Prior*Size	.000	.004	.000			
	Prior*Location	.000	.052	.016			
	Prior*Size*Sample size	.003	.006	.021			
	Prior*Location*Sample size	.026	.003	.030			
	Prior*Size*Location	.003	.012	.002			
Prior*Size*Location*Sample size	.004	.009	.001				

S- $X^2$ . Using S- $X^2$  as the discrepancy statistic, the interaction term of the two-way ANOVA was not significant ( $F(4, 2907) = 0.60, p = .6659$ ) between sample size and

prior informativeness. Significant main effects were observed for sample size ( $F(2, 2907) = 116.47, p < .0001$ ), but not prior informativeness ( $F(2, 2907) = 0.28, p = .7582$ ). The main effect of sample size showed significant mean hit rate differences between 500 people ( $M_{hit} = 0.20$ ) and 1000 people ( $M_{hit} = 0.35, M_{hit\_diff} = 0.15, p < .0001$ ) and 2500 people ( $M_{hit} = 0.52, M_{hit\_diff} = 0.32, p < .0001$ ). Significant mean hit rate differences also occurred between 1000 people and 2500 people ( $M_{hit\_diff} = 0.17, p < .0001$ ).

**INFIT.** Using *INFIT* as the discrepancy statistic, the interaction term of the two-way ANOVA was not significant ( $F(4, 2907) = 1.34, p = .2517$ ) between sample size and prior informativeness. Significant main effects were observed for sample size ( $F(2, 2907) = 31.50, p < .0001$ ) and prior informativeness ( $F(2, 2907) = 98.42, p < .0001$ ). The main effect of sample size showed significant mean hit rate differences between 500 people ( $M_{hit} = 0.53$ ) and 2500 people ( $M_{hit} = 0.70, M_{hit\_diff} = 0.17, p < .0001$ ). Significant mean hit rate differences also occurred between 1000 people ( $M_{hit} = 0.58$ ) and 2500 people ( $M_{hit\_diff} = 0.12, p < .0001$ ).

Mean hit rates were significantly different for noninformative priors ( $M_{hit} = 0.44$ ), informative-accurate priors ( $M_{hit} = 0.63$ ) and informative-inaccurate priors ( $M_{hit} = 0.73$ ). Specifically, the mean hit rate for noninformative priors was found to be significantly lower than the mean hit rate for informative-accurate priors ( $M_{hit\_diff} = 0.19$ ) and significantly lower than the mean hit rate for informative-inaccurate priors ( $M_{hit\_diff} = 0.29$ ). Further, the hit rate for informative-accurate priors was found to be significantly lower than the hit rate for informative-inaccurate priors ( $M_{hit\_diff} = 0.10$ ).

**Percent correct.** Using percent correct as the discrepancy statistic, the interaction term of the two-way ANOVA was significant ( $F(8, 2907) = 21.06, p < .0001$ ) between sample size and prior informativeness. Within informative-accurate priors, the effect of sample size was significant, with the mean hit rate for 500 people significantly lower than the mean hit rate for 2500 people ( $M_{hit\_diff} = 0.15, p < .0001$ ) and the mean hit rate for 1000 people significantly lower than the hit rate for 2500 people ( $M_{hit\_diff} = 0.12, p < .0001$ ). Within informative-inaccurate priors and noninformative priors, on the other hand, there were no significant mean differences between hit rates for the three sample sizes. The mean hit rates, when using percent correct as the discrepancy statistic, were lower than those found with the other two discrepancy statistics.

### **False Positive Error**

Table 13 provides a summary of the mean false positive rates for Research Question 2 and Figure 11 presents the results graphically.

Table 13. Mean False Positive Rates by Prior Specification and Sample size

Prior	Sample size	$S-X^2$	<i>INFIT</i>	Percent correct***
Informative- Accurate	500	.17	.17	.001
	1000	.19	.21	.01
	2500	.20	.33	.05
Informative- Inaccurate	500	.17	.30	.01
	1000	.18	.31	.01
	2500	.21	.46	.09
Noninformative	500	.16	.12	.000
	1000	.18	.16	.01
	2500	.21	.28	.03

*Note*, \*\*\* indicates statistical significance of the two-way ANOVA at the .001 level.

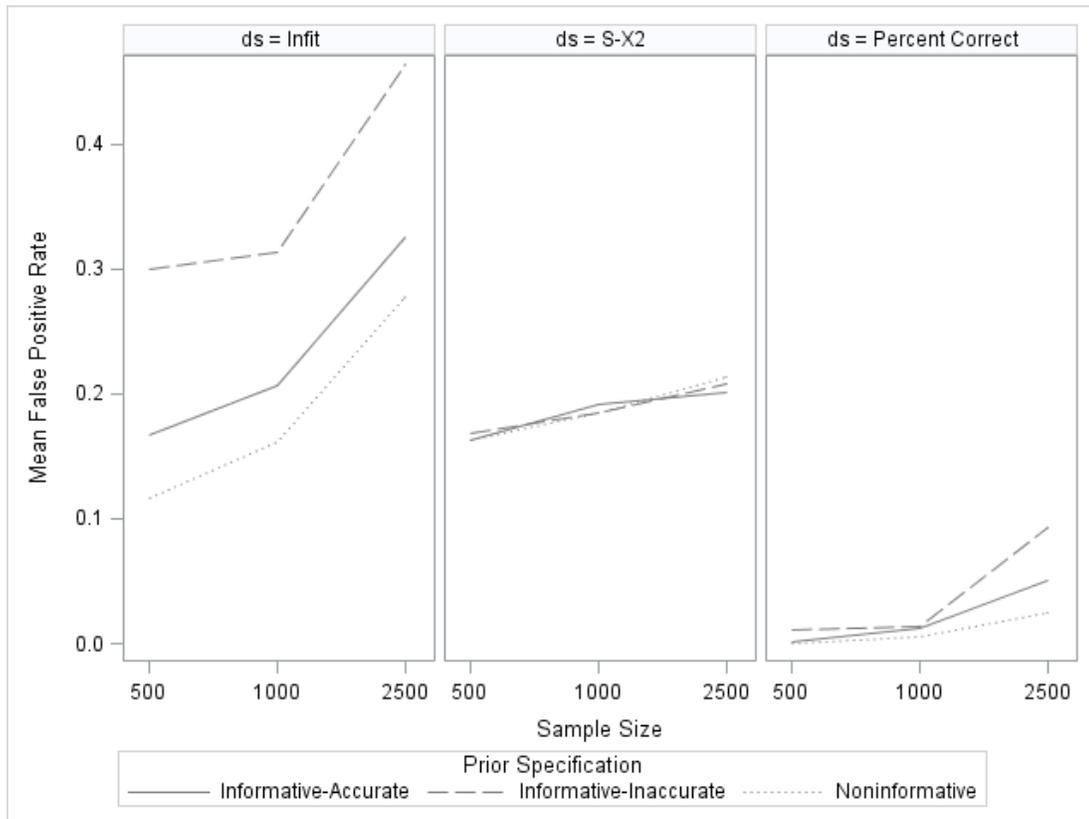


Figure 11. Mean False Positive Rates, by Type of Prior, Sample size, and Discrepancy Statistic

**S-X<sup>2</sup>.** Using  $S-X^2$  as the discrepancy statistic, the interaction term of the two-way ANOVA was not significant ( $F(4, 10035) = 0.23, p = .9201$ ) between sample size and prior informativeness for false positive error rates. Significant main effects were observed for sample size ( $F(2, 10035) = 9.95, p < .0001$ ) but not prior informativeness ( $F(2, 10035) = 0.02, p < .0001$ ). The main effect of sample size showed significant mean false positive error rate differences between 500 people ( $M_{false+} = 0.16$ ) and 2500 people ( $M_{false+} = 0.21, M_{false+\_diff} = 0.05, p < .0001$ ). As with hit rates, as sample size increased, so too did the false positive error rates, within levels of prior informativeness.

**INFIT.** Using *INFIT* as the discrepancy statistic, the two-way ANOVA was not significant ( $F(4, 10035) = 0.64, p = .6365$ ) between sample size and prior informativeness for false positive error rates. Significant main effects were observed for sample size ( $F(2, 10035) = 134.85, p < .0001$ ) and prior informativeness ( $F(2, 10035) = 148.52, p < .0001$ ). The main effect of sample size showed significant mean false positive error rate differences between 500 people ( $M_{false+} = 0.19$ ) and 2500 people ( $M_{false+} = 0.36, M_{false+_{diff}} = 0.17, p < .0001$ ). Significant mean false positive error rate differences also occurred between 1000 people ( $M_{false+} = 0.23$ ) and 2500 people ( $M_{false+_{diff}} = 0.13, p < .0001$ ).

Mean false positive error rates were significantly different for noninformative priors ( $M_{false+} = 0.19$ ), informative-accurate priors ( $M_{false+} = 0.23$ ) and informative-inaccurate priors ( $M_{false+} = 0.34$ ) - specifically, the false positive error rate for noninformative priors was found to be significantly lower than the false positive error rate for informative-accurate priors ( $M_{false+_{diff}} = 0.04$ ) and significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.15$ ). Further, the false positive error rate for informative-accurate priors was found to be significantly lower than the false positive error rate for informative-inaccurate priors ( $M_{false+_{diff}} = 0.11$ ). The false positive error rates were all greater than the desirable level of .10. As with hit rates, as the sample size increased, so too did the false positive error rates within the prior specification.

**Percent correct.** Using percent correct as the discrepancy statistic, the interaction term of the two-way ANOVA was significant ( $F(8, 10035) = 21.06, p < .0001$ ) between

sample size and prior informativeness for false positive error rates. Within informative-accurate priors, the effect of sample size was significant, with the false positive error rate for 500 people ( $M_{false+} = 0.002$ ) significantly lower than the false positive error rate rate for 2500 people ( $M_{false+} = 0.05$ ,  $M_{false+_{diff}} = 0.048$ ,  $p < .0001$ ) and the false positive error rate for 1000 people ( $M_{false+} = 0.01$ ) significantly lower than the false positive error rate for 2500 people ( $M_{false+_{diff}} = 0.049$ ,  $p < .0001$ ). Within informative-inaccurate priors, there were no significant mean differences between hit rates for the three sample sizes. Within noninformative priors, the false positive error rate for 1000 people ( $M_{false+} = 0.005$ ) was significantly lower than the false positive error rate for 2500 people ( $M_{false+} = 0.03$ ,  $M_{false+_{diff}} = 0.025$ ,  $p < .0001$ ). False positive error rates for the percent correct discrepancy statistic were much lower than for the other discrepancy statistics.

### **Summary**

Hit rates and false positive error rates showed a significant interaction between sample size and prior informativeness when using percent correct as a discrepancy statistic. For all three discrepancy statistics, the main effect of sample size was significant. As sample size increased, hit rates and false positive error rates also increased, within prior specification and across discrepancy statistics.

### **Research Question 3: How Does the Type of Misfit Affect the Influence of Prior Specification on the Results of the PPC Method for the Model-data Fit of Unidimensional, Dichotomous IRT Models?**

Two types of misfit were investigated: size of misfit and type of misfit. Size of misfit had two levels, small and medium-large, as described in Chapter 3. Type of misfit

refers to the location along the latent ability continuum where the largest difference between the GM and AM occurs. For example, if the GM is the 3PL and the AM is the 2PL or 1PL, the largest degree of misfit occurs at the lower end of the ability continuum. If the GM is the LPE and the AM is the 3PL, 2PL, or 1PL, the largest degree of misfit occurs at the higher end of the ability continuum. If the GM is the 2PL and the AM is the 1PL, the type of misfit relates to the steepness of the slope of the IRFs.

Hit rates only were considered for this research question, as all items generated to have adequate model-data fit would have only a small degree of misfit. False positive error rates would not have sufficiently varying levels of size of misfit. A fully-crossed, three-way ANOVA, with location of misfit, size of misfit, and informativeness of prior distributions as factors was conducted for this research question.

### **Hit Rates**

Table 14 provides a summary of the mean hit rates used to answer Research Question 3 and Figure 12 illustrates the hit rate means.

Table 14. Mean Hit Rates by Prior Specification, Sample size, and Type if Misfit

Prior	Type	Size	SX2	<i>INFIT</i> ***	Percent correct
Informative-Accurate	lower	Medium-Large	.42	.40	.01
		Small	.06	.11	.06
	slope	Medium-Large	.57	.92	.00
		Small	.22	1.00	.00
	upper	Medium-Large	.29	.69	.14
		Small	.00	1.00	.08
Informative-Inaccurate	lower	Medium-Large	.42	.67	.00
		Small	.00	.67	.00
	slope	Medium-Large	.58	.93	.01
		Small	.33	.89	.00
	upper	Medium-Large	.32	.72	.07
		Small	.00	.67	.00
Noninformative	lower	Medium-Large	.42	.43	.01
		Small	.00	.06	.00
	slope	Medium-Large	.54	.94	.00
		Small	.44	.89	.00
	upper	Medium-Large	.30	.32	.02
		Small	.00	.00	.00

*Note*, \*\*\* indicates statistical significance of the three-way ANOVA at the .001 level.

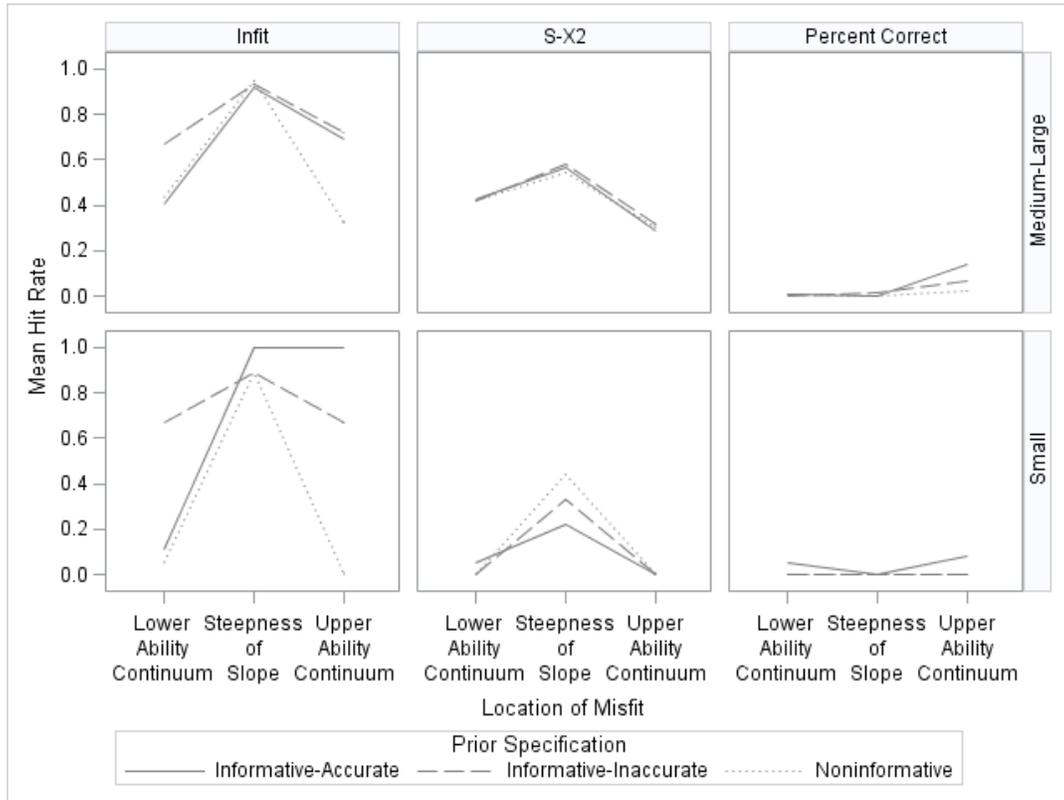


Figure 12. Mean False Positive Rates, by Type of Prior, Sample size, and Discrepancy Statistic

**S-X<sup>2</sup>.** Using  $S-X^2$  as the discrepancy statistic, the main effects of size of misfit ( $F(1, 2898) = 88.44, p < .0001$ ) and location of misfit ( $F(2, 2898) = 25.02, p < .0001$ ) were significant. There were no significant two-way or three-way interactions. The main effect of size of misfit showed significant hit rate mean differences between small misfit ( $M_{hit} = 0.22$ ) and large misfit ( $M_{hit} = 0.43, M_{hit\_diff} = 0.21, p < .0001$ ).

The main effect of location of misfit showed significant mean hit rate differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.22$ ) and misfit related to steepness of the IRF slope ( $M_{hit} = 0.45, M_{hit\_diff} = 0.23, p < .0001$ ). There was also a significant mean hit rate difference between misfit at the upper end of the ability

continuum ( $M_{hit} = 0.15$ ) and misfit related to steepness of the IRF slope ( $M_{hit\_diff} = 0.30$ ,  $p < .0001$ ). The PPC procedure had almost no ability to detect small degrees of misfit when the misfit was located at the upper end of the ability continuum.

**INFIT.** Using *INFIT* as the discrepancy statistic, the three-way interaction was significant ( $F(6, 2898) = 5.43$ ,  $p < .0001$ ) between size of misfit, location of misfit, and prior informativeness. Within informative-accurate priors, for items with large misfit, there were significant mean differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.40$ ) and the upper end of the ability continuum ( $M_{hit} = 0.69$ ,  $M_{hit\_diff} = 0.29$ ,  $p < .0001$ ) and slope steepness ( $M_{hit} = 0.92$ ,  $M_{hit\_diff} = 0.52$ ,  $p < .0001$ ). There were also significant mean hit rate differences between misfit at the upper end of the ability continuum and misfit related to slope steepness ( $M_{hit\_diff} = 0.23$ ). Within informative-accurate priors, for items with small misfit, there were significant mean differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.11$ ) and the upper end of the ability continuum ( $M_{hit} = 0.99$ ,  $M_{hit\_diff} = 0.88$ ,  $p < .0001$ ) and slope steepness ( $M_{hit} = 0.99$ ,  $M_{hit\_diff} = 0.88$ ,  $p < .0001$ ).

Within informative-inaccurate priors, for items with large misfit, there were significant mean differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.67$ ) and slope steepness ( $M_{hit} = 0.93$ ,  $M_{hit\_diff} = 0.27$ ,  $p < .0001$ ). There were also significant mean hit rate differences between misfit at the upper end of the ability continuum ( $M_{hit} = 0.72$ ) and misfit related to slope steepness ( $M_{hit\_diff} = 0.21$ ,  $p < .0001$ ). Within informative-inaccurate priors, for items with small misfit, there were no significant mean differences between misfit at the lower end of the ability continuum

( $M_{hit} = 0.67$ ), the upper end of the ability continuum ( $M_{hit} = 0.67$ ), or slope steepness ( $M_{hit} = 0.89$ ).

Within noninformative priors, for items with large misfit, there were significant mean differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.43$ ) and slope steepness ( $M_{hit} = 0.94$ ,  $M_{hit\_diff} = 0.51$ ,  $p < .0001$ ). There were also significant mean hit rate differences between misfit at the upper end of the ability continuum ( $M_{hit} = 0.32$ ) and misfit related to slope steepness ( $M_{hit\_diff} = 0.52$ ,  $p < .0001$ ). Within noninformative priors, for items with small misfit, there were significant mean differences between misfit at the lower end of the ability continuum ( $M_{hit} = 0.06$ ) and slope steepness ( $M_{hit} = 0.89$ ,  $M_{hit\_diff} = 0.83$ ,  $p < .0001$ ). There were also significant mean hit rate differences between misfit at the upper end of the ability continuum ( $M_{hit} = 0.02$ ) and misfit related to slope steepness ( $M_{hit\_diff} = 0.87$ ,  $p < .0001$ ).

In Figure 12, the left panel provides a comparison of results for *INFIT*. It appears *INFIT* performs best when the misfit is large or related to slope steepness. As with the *S-X<sup>2</sup>* results, when the misfit is small or at the upper end of the ability continuum, the PPC procedure does a poor job of detecting misfit.

**Percent correct.** Using percent correct as a discrepancy statistic, the three-way ANOVA had no significant main effects or interactions. This is likely due to the very low hit rates across all conditions when using percent correct as the discrepancy statistic.

### Summary

A significant mean difference in hit rates occurred for the main effects of size and location when using *S-X<sup>2</sup>* as the discrepancy statistic. A significant three-way interaction

occurred between prior informativeness, size, and location of misfit when using *INFIT* as the discrepancy statistic. The PPC procedure performed best when misfit was large and related to steepness of slope.

**Research Question 4: How Does the Interaction of Sample Size and Type of Misfit Affect the Influence of Prior Specification on the Results of the PPC Method for the Model-data Fit of Unidimensional, Dichotomous IRT Models?**

Following the same approach as was used for Research Question 3, two types of misfit were investigated: size of misfit and type of misfit. A fully-crossed, four-way ANOVA, with type of misfit, size of misfit, sample size, and informativeness of prior distributions was conducted for this research question. Hit rates only were considered for this research question, as all items generated to have adequate fit would have only a small degree of misfit.

**Hit Rates**

*S-X<sup>2</sup>*. Using *S-X<sup>2</sup>* as the discrepancy statistic, there were no significant two-way interactions, three-way interactions, or four-way interactions. The main effects of size of misfit ( $F(1, 2862) = 98.15, p < .0001$ ), location of misfit ( $F(2, 2862) = 73.12, p < .0001$ ), and number of people ( $F(2, 2862) = 128.57, p < .0001$ ) were significant.

*INFIT*. Using *INFIT* as the discrepancy statistic, the four-way interaction was not significant ( $F(12, 2862) = 0.27, p = .0090$ ) between size of misfit, location of misfit, sample size, and prior informativeness. The three-way interaction was significant, as it was for Research Question 3.

**Percent correct.** Using percent correct as the discrepancy statistic, the four-way interaction was not significant ( $F(12, 2862) = 2.22, p = .9938$ ) between size of misfit, location of misfit, sample size, and prior informativeness. The three-way interaction was significant, as it was for Research Question 3.

### **Summary**

None of the three discrepancy statistics indicated a significant four-way interaction between size of misfit, location of misfit, sample size, and prior informativeness. This indicates that the effect of sample size on prior specification and the PPC procedure is not influenced by type of misfit present.

### **Other Considerations**

**Percent of misfitting items.** The percent of misfitting items was also considered as a factor for all conditions. When used in a two-way ANOVA for hit rate, none of the discrepancy statistics showed significant mean differences for the main effect of percent of misfit or for the two-way interaction of percent of misfit and informativeness of prior distributions.

When used in a two-way ANOVA for false positive error rates, there was a significant interaction between prior informativeness and percent of misfitting items ( $F(6, 10032) = 12.77, p < .0001$ ) with *INFIT* used as the discrepancy statistic. The same significant interaction is not present when  $S-X^2$  or percent correct are used as the discrepancy statistic. These results are presented in Table 15 and Figure 13.

Table 15. Mean False Positive Rates by Type of Prior and Percent of Misfitting Items, for INFIT

Prior	Percent Misfitting	<i>INFIT</i>
Informative-Accurate	0%	0.09
	10%	0.19
	30%	0.38
	50%	0.40
Informative-Inaccurate	0%	0.16
	10%	0.29
	30%	0.57
	50%	0.58
Noninformative	0%	0.09
	10%	0.14
	30%	0.30
	50%	0.29

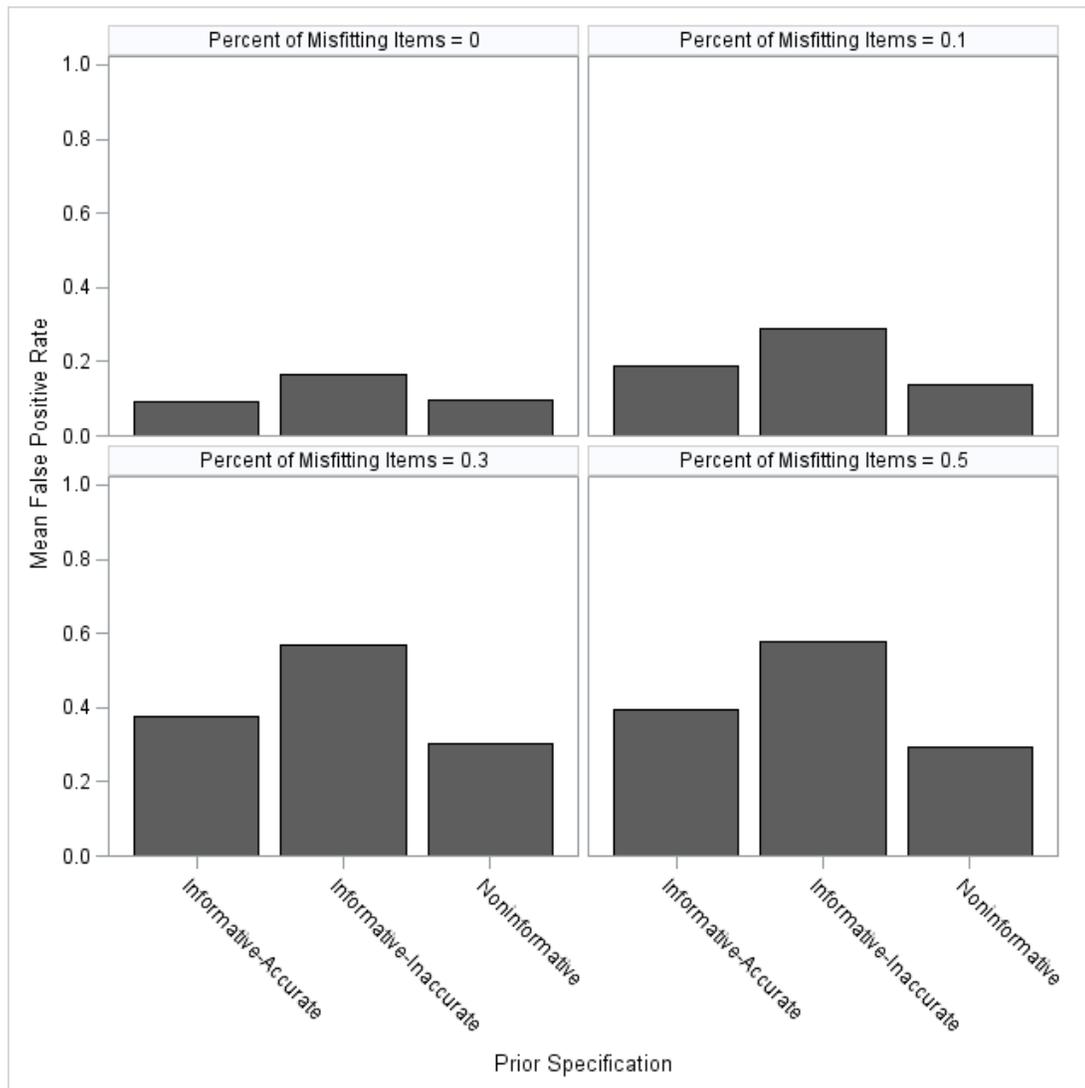


Figure 13. Mean False Positive Rates, by Type of Prior, Percent of Misfit, for INFIT

Specifically, within informative-accurate priors, there were significant mean differences for false positive error rates between 0% misfitting ( $M_{false+} = 0.09$ ) and 10% misfitting ( $M_{false+} = 0.19$ ,  $M_{false+\_diff} = 0.10$ ,  $p < .0001$ ), between 0% misfitting and 30% misfitting ( $M_{false+} = 0.38$ ,  $M_{false+\_diff} = 0.29$ ,  $p < .0001$ ), and between 0% misfitting and 50% misfitting ( $M_{false+} = 0.40$ ,  $M_{false+\_diff} = 0.31$ ,  $p < .0001$ ). There also were significant

mean differences for false positive error rates between 10% misfitting ( $M_{false+} = 0.09$ ) and 30% misfitting ( $M_{false+_{diff}} = 0.19, p < .0001$ ), and between 10% misfitting and 50% misfitting ( $M_{false+_{diff}} = 0.21, p < .0001$ ).

Within informative-inaccurate priors, there were significant mean differences for false positive error rates between 0% misfitting ( $M_{false+} = 0.16$ ) and 30% misfitting ( $M_{false+} = 0.57, M_{false+_{diff}} = 0.41, p < .0001$ ), and between 0% misfitting and 50% misfitting ( $M_{false+} = 0.58, M_{false+_{diff}} = 0.42, p < .0001$ ). There also were significant mean differences for false positive error rates between 10% misfitting ( $M_{false+} = 0.29$ ) and 30% misfitting ( $M_{false+_{diff}} = 0.29, p < .0001$ ), and between 10% misfitting and 50% misfitting ( $M_{false+_{diff}} = 0.29, p < .0001$ ).

Within noninformative priors, there were significant mean differences for false positive error rates between 0% misfitting ( $M_{false+} = 0.09$ ) and 30% misfitting ( $M_{false+} = 0.30, M_{false+_{diff}} = 0.21, p < .0001$ ), and between 0% misfitting and 50% misfitting ( $M_{false+} = 0.29, M_{false+_{diff}} = 0.20, p < .0001$ ). There also were significant mean differences for false positive error rates between 10% misfitting ( $M_{false+} = 0.14$ ) and 30% misfitting ( $M_{false+_{diff}} = 0.16, p < .0001$ ), and between 10% misfitting and 50% misfitting ( $M_{false+_{diff}} = 0.15, p < .0001$ ). As the percent of misfitting items increased, so too did the false positive error rates.

Figures 14 through 19 show the differences in hit rates for the combination of percent misfitting items and size of misfit as well as the combination of percent misfitting items and location of misfit. It can be seen in these results that choice of discrepancy statistic influences the hit rates. When *INFIT* is used, and misfit is small or large,

informative-inaccurate priors had the highest hit rates, except in the case of 500 people and 50% misfitting items. Noninformative priors always had the lowest mean hit rates when *INFIT* was used. There were no items with small misfit for conditions with 10% misfitting items, which is why the lower left panel is blank in Figure 14.

In Figure 15, *INFIT* is still the discrepancy statistic, but the patterns regarding location of misfit are not as clear. What is evident is that for misfit related to steepness of the slope (for instance, when GM=2PL, AM=1PL), the PPC procedure performs best, with hit rates as high as 0.99, regardless of prior specification used. In general, noninformative priors had the lowest hit rates, regardless of the location of the misfit.

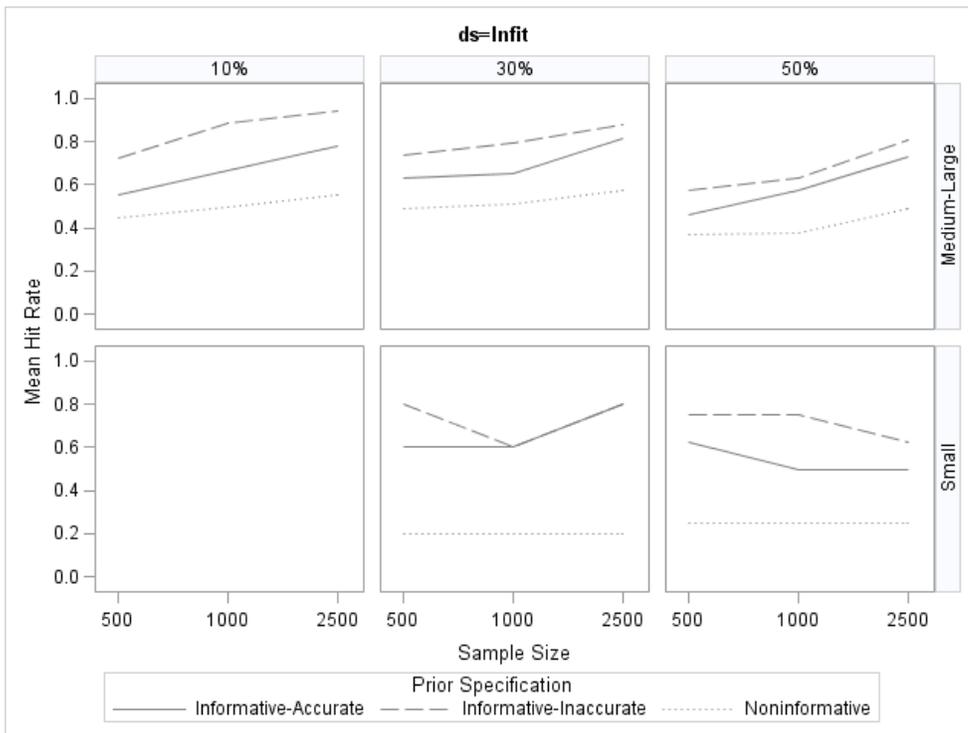


Figure 14. Hit Rates, with *INFIT* as Discrepancy Statistic, by Size of Misfit and Percent Misfit

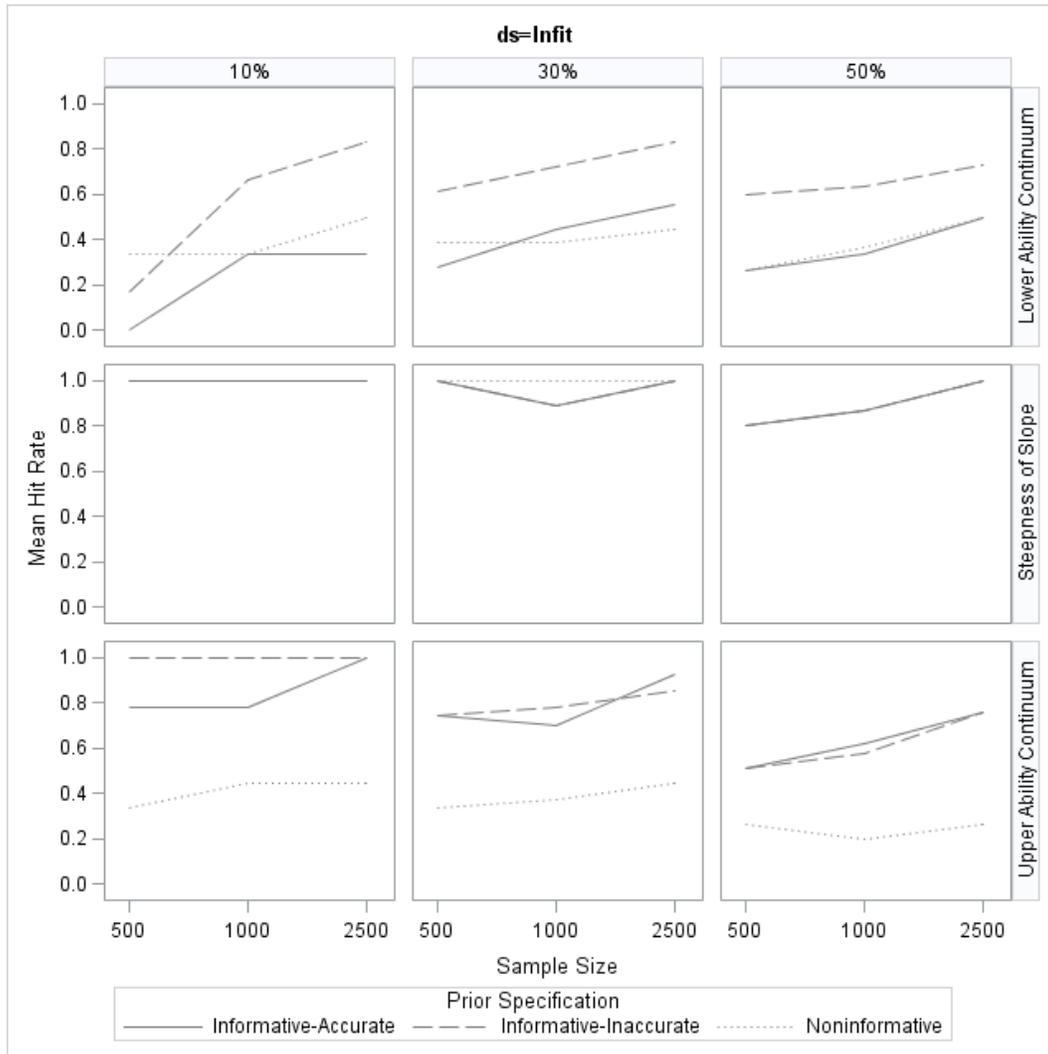


Figure 15. Hit Rates, with INFIT as Discrepancy Statistic, by Location of Misfit and Percent Misfit

Figures 16 and 17 illustrate that, when  $S-X^2$  is used as the discrepancy statistic, there is little divergence in hit rates for each type of prior used. One exception is the case of 50% misfitting items and small misfit and 1000 people - in this scenario, informative-accurate priors tend to do best and noninformative priors perform the worst. Another exception is 50% misfitting items, misfit related to slope steepness, and 1000 people, in

which informative-accurate priors also tend to do best and noninformative the worst.

There were no items with small misfit for conditions with 10% misfitting items, which is why the lower left panel is blank in Figure 16.

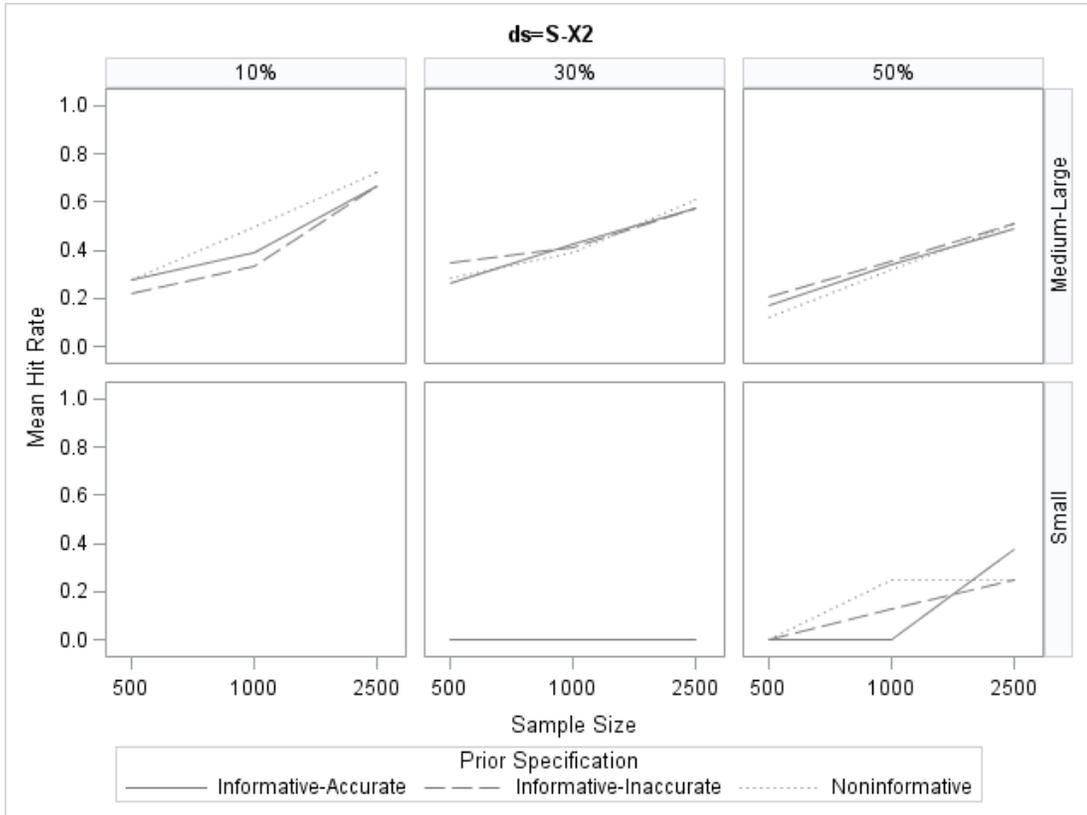


Figure 16. Hit Rates, with  $S-X^2$  as Discrepancy Statistic, by Size of Misfit and Percent Misfit

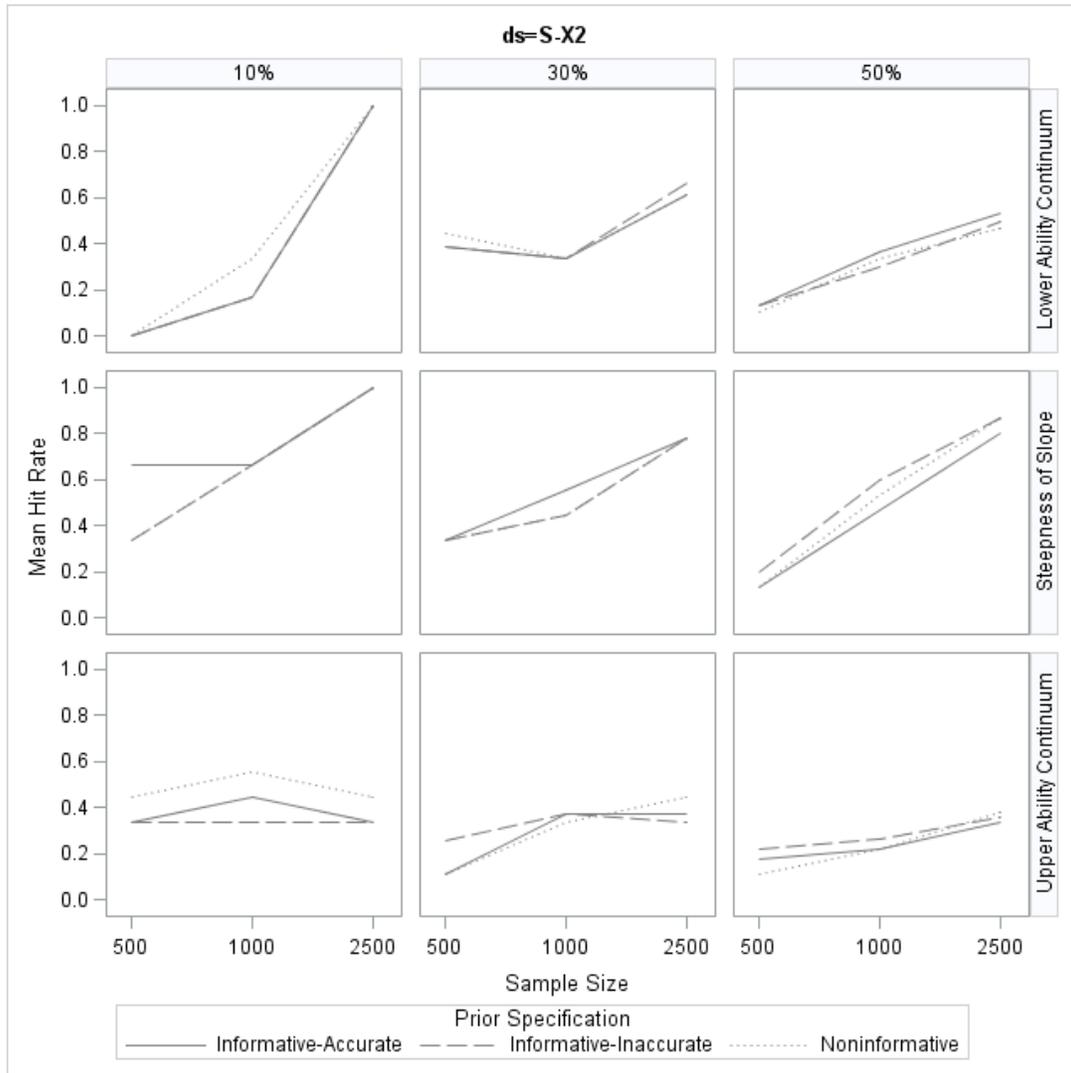


Figure 17. Hit Rates, with  $S-X^2$  as Discrepancy Statistic, by Location of Misfit and Percent Misfit

Figures 18 and 19 show the patterns related to prior informativeness with percent correct as the discrepancy statistic, but the hit rates are so low that the discrepancy statistic is deemed inadequate in all scenarios. There were no items with small misfit for conditions with 10% misfitting items, which is why the lower left panel is blank in Figure 18.

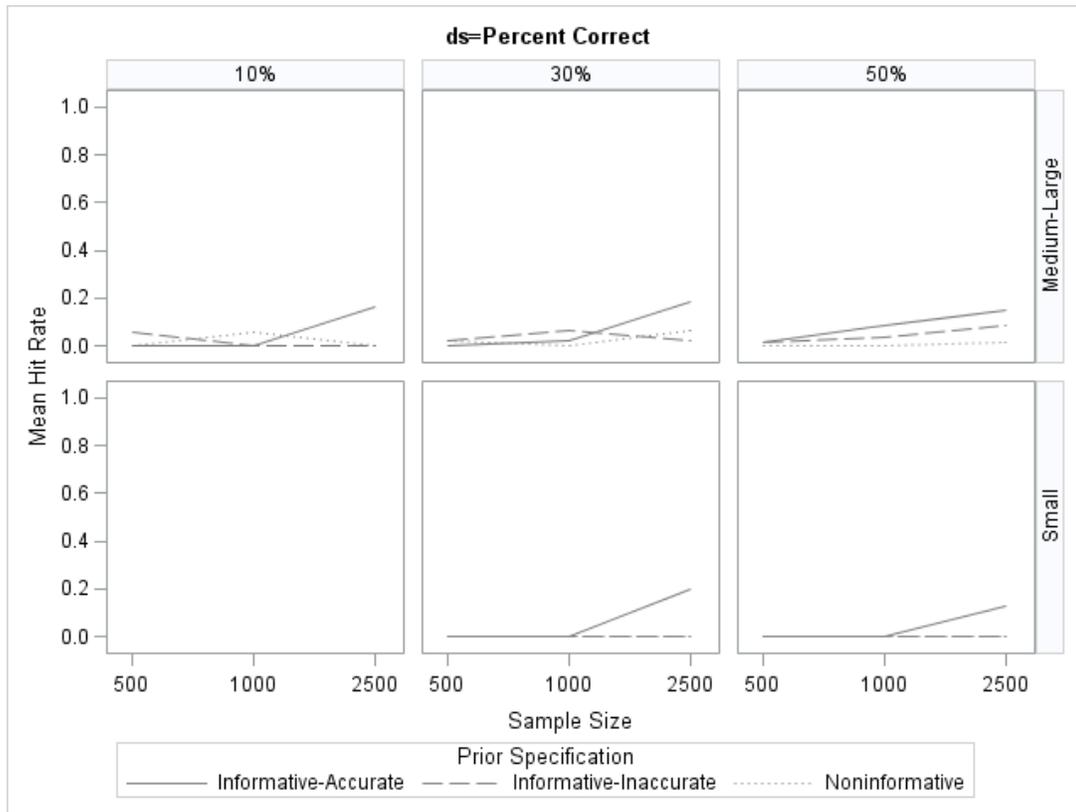


Figure 18. Hit Rates, with Percent correct as Discrepancy Statistic, by Size of Misfit and Percent Misfit

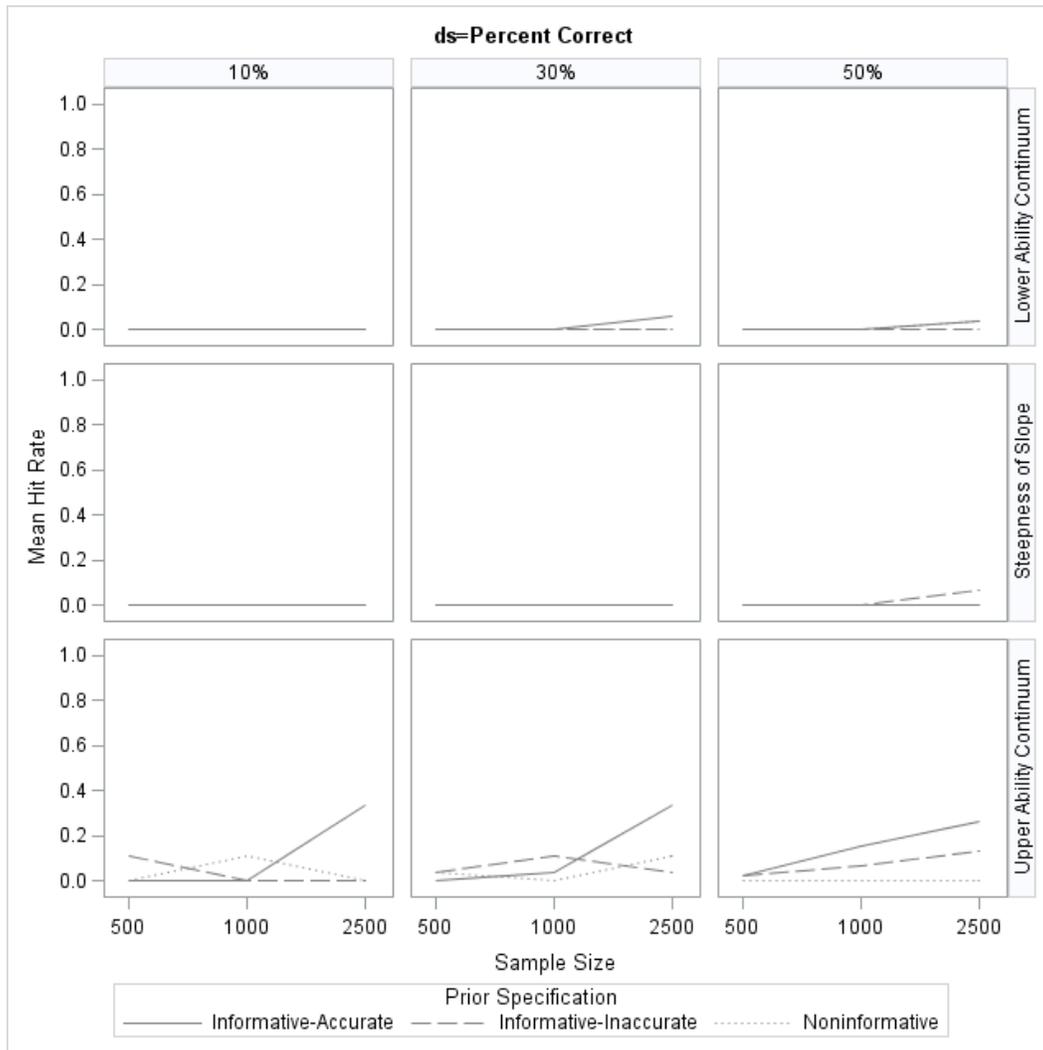


Figure 19. Hit Rates, with Percent correct as Discrepancy Statistic, by Location of Misfit and Percent Misfit

### Summary

In this chapter, I have presented the results of the simulation study used to address the four research questions. The conditions tested in the study design included sample size (500, 1000, 2500), percent of misfitting items (0%, 10%, 30%, 50%), degree of

misfit (small, medium-large), and type of misfit introduced when the GM and AM were different.

Geweke diagnostics and trace plots were used to assess convergence. The number of burn-in iterations and total iterations had to be increased to achieve convergence of the Markov chain for parameter posterior distributions. The final overall number of iterations used was 5,000 for burn-in with 12,000 iterations following the burn-in. This was adequate to achieve convergence in one replication of each scenario for the cases when priors were specified as informative-accurate, informative-inaccurate, and noninformative. When the prior specification was noninformative-inaccurate, the proposal distributions were not fully tuned and convergence could not be assessed. At this stage, model-data fit would not be able to be assessed, consequently this prior specification was not included in subsequent analyses.

For the prior specifications that did converge, informative-accurate priors resulted in faster convergence, on average. When the percent of misfitting items was low (0% or 10% misfitting), noninformative priors tended to result in slightly faster convergence times than those of informative-inaccurate priors. However, when the percent of misfitting items was high (30% or 50% misfitting), noninformative priors tended to result in slower convergence times than those of informative-inaccurate priors. Convergence times, compared across priors, were closest together with smaller sample sizes, and farther apart for larger sample sizes. In all cases, time to converge increased as sample size increased.

Each of three discrepancy statistics ( $S-X^2$ , *INFIT*, and percent correct) were evaluated separately for false positive error rates and hit rates. Fully crossed ANOVAs were performed to investigate mean differences among hit rates and false positive error rates, with the significance level for the ANOVA tests and follow-up tests set to .001.

For the one-way ANOVA, investigating only the main effects of prior specification, the discrepancy statistics of *INFIT* and percent correct showed significant mean differences for hit rate and false positive error rate. Informative-inaccurate priors had the highest hit and false positive error rates for these discrepancy statistics. This indicates that the use of informative-inaccurate priors, and, in particular, *INFIT* as the discrepancy statistic, tends to flag most items as misfitting. Use of percent correct had the lowest false positive error rates across all prior specifications, but the hit rate was unacceptably low for this discrepancy statistic.

Two-way ANOVAs indicated that as sample size increased, so too did hit rates and false positive error rates, regardless of which prior specification was used. However, when informative-inaccurate priors were used, and *INFIT* was the discrepancy statistic, false positive error rates were greater than not only the .05 traditionally desirable level, but also the false positive error rates of the other prior specifications.

As the size of the misfit increased from small to medium-large, so too did the hit rates for all three discrepancy statistics. When the location of misfit was investigated, misfit related to the steepness of the IRF slope tended to show the highest hit rates when  $S-X^2$  and *INFIT* were used. None of the three discrepancy statistics indicated a significant four-way interaction between size of misfit, location of misfit, sample size, and prior

informativeness. As the percent of misfitting items increased, so too did false positive error rates, but only when *INFIT* was used as the discrepancy statistic.

## CHAPTER V

### DISCUSSION AND IMPLICATIONS

#### **Conclusions**

First, a brief summary is presented of the findings for each of the four research questions.

**Research question 1: Prior specification on PPC results.** Fully crossed ANOVAs were performed to investigate mean differences among hit rates and false positive error rates. There were differences in both hit rates and false positive error rates, for the discrepancy statistics of *INFIT* and percent correct, depending on prior specification. Informative-inaccurate priors had the highest hit and false positive error rates for these two types of discrepancy statistics. However, the use of percent correct was a poor choice for discrepancy statistic, since, while the false positive error rate tended to be very low for this discrepancy statistic, it also had very low hit rates. The use of informative-inaccurate priors, and, in particular, with *INFIT* as the discrepancy statistic, tended to classify most items as misfitting, resulting in elevated false positive error rates.

**Research question 2: Prior specification, sample size on PPC results.** In general, as sample size increased, so too did hit rates and false positive error rates, regardless of which prior specification was used. However, when informative-inaccurate priors were used, and *INFIT* was the discrepancy statistic, false positive error rates were

greater than not only the .05 traditionally desirable level, but also the false positive error rates of the other prior specifications.

**Research question 3: Prior specification, type of misfit on PPC results.** As size of misfit increased, the hit rates for all three discrepancy statistics also increased. When the location of misfit was investigated, misfit related to the steepness of the IRF slope tended to show the highest hit rates when  $S-X^2$  and *INFIT* were used. As percent of misfitting items increased, so too did false positive error rates, but only when *INFIT* was used as the discrepancy statistic.

**Research question 4: Prior specification, sample size, type of misfit on PPC results.** None of the three discrepancy statistics indicated a significant four-way interaction between size of misfit, location of misfit, sample size, and prior informativeness.

### **Implications**

The PPC procedure hit rates appear to be influenced by prior specification, but only in some instances. The effect of prior informativeness is tied to the choice of discrepancy measure used. For instance, returning to Figure 9, when *INFIT* is used as the discrepancy statistic, hit rates for informative-inaccurate priors were higher than for informative-accurate priors, with noninformative priors having the lowest hit rates. When  $S-X^2$  or percent correct were used, the effect of prior was negligible. Previous research in PPC for IRT has varied the discrepancy statistic, while keeping the prior unchanged within a study, which likely masked these types of differences.

The PPC procedure false positive rate also appears to be influenced by prior specification. Again, this is not a blanket statement and is applicable only in some instances. For hit rates, the effect of prior informativeness is also tied to the choice of discrepancy measure used. Again, with *INFIT* used as the discrepancy statistic, false positive rates for informative-inaccurate priors were higher than for informative-accurate priors and noninformative priors. When  $S-X^2$  or percent correct were used, the effect of prior was negligible on false positive rates.

In general, the use of *INFIT* and  $S-X^2$  produced very high false positive rates. For example, mean false positive rates, found in Table 11, are as high as .46 (*INFIT* as the discrepancy statistic, 2500 individuals, and informative-inaccurate priors). This is well above the nominal .10 error rate for the PPC method. Regarding informative-accurate and noninformative priors, the false positive rates were similar to what some other studies investigating PPC for IRT have found. For instance, Toribio and Albert (2011) found false positive rates as high as .16 for detecting misfit of unidimensional IRT models. Glas and Meijer (2003), for tests with 30 items, some person-fit discrepancy measures saw error rates of .18 and .19, and even higher for longer tests (as high as .26). Sinharay (2005) found much lower error rates (below .05) for detecting misfit of unidimensional IRT models. However, this study used different thresholds for extreme PPP values. Rather than  $PPP < .05$  or  $PPP > .95$ , Sinharay (2005) used  $PPP < .025$  or  $PPP > .975$  to flag an item as misfitting, which effectively makes the PPC method more conservative.

Changing the PPP extreme values from  $PPP < .05$  or  $PPP > .95$  to  $PPP < .025$  or  $PPP > .975$ , had a considerable impact for informative-inaccurate priors. The percent of

items flagged as misfitting decreased by 14% as a result of this change when *INFIT* is used as the discrepancy statistic. When  $S-X^2$  is used as the discrepancy statistic, for informative-inaccurate priors, changing the PPP extreme value resulted in 32% fewer items flagged as misfitting. And when percent correct is used as the discrepancy, the same change in extreme PPP values saw a 42% decrease in items flagged as misfitting. This small adjustment to the PPC procedure provides a significant change to the outcome and should be given more attention in future studies.

### **Item and Person Location**

*INFIT* is more sensitive to prior specification than  $S-X^2$  due to the method of computation for the statistics. As described in Chapter 2, *INFIT* is an information-weighted sum (Bond & Fox, 2001). The statistical information in an item is related to its variance, which is larger for well-targeted observations and smaller for extreme observations. In other words, *INFIT* is sensitive to unexpected responses near the person's ability estimate.

The data were generated with random person ability parameters, drawn from a standard normal distribution ( $\theta \sim N(0, 1)$ ). As seen in Tables 4, 5, and 6, item difficulty values on misfitting items were never made more extreme than item difficulty values on items with adequate model-data fit. Item 2 in Table 5 is an example of this - when the item is designed to fit, the item's difficulty value is -1.9, 1.9 standard deviations below the center of the generating person ability distribution. In the conditions where item 2 is designed to contain misfit (i.e., 30% misfitting items or 50% misfitting items), the item's

difficulty value is -0.21, which is still below the center of the source person ability distribution, but less than one quarter of a standard deviation away.

Item difficulty values inadvertently were moved closer to the center of the generating person ability distribution. Table 16 illustrates the change in center and standard deviation of the difficulty values. Although the mean difficulty value remained relatively consistent, close to zero, the spread of values decreased as misfit increased. Since there are more values close to the center of the ability distribution, *INFIT* is at an advantage in that it is better able to detect misfit in this region.

Table 16. Mean Item Difficulty Values

	0% Misfitting	10% Misfitting	30% Misfitting	50% Misfitting
Mean	0	-0.029	0.015	0.175
Standard Deviation	1.2135	1.159	1.066	0.999

### **Type of Misfit**

Items with misfit related to steepness of slope were also detected with a higher hit rate than other forms of misfit. Ideally, test statistics should be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied (Gelman et al., 2003, p. 172; Sinharay, 2005). Use of  $S-X^2$  to detect misfit at the lower end of the ability continuum has been investigated by Orlando and Thissen (2005). They found that when the generating model was the 3PL and the analysis model (what they refer to as the “calibrating model”) is the 1PL, the estimated IRF underestimated the proportion of correct responses at the low end of the latent ability continuum, where there

were fewer observations. This, they concluded, made it more difficult to detect misfit associated with the lower end of the latent ability continuum.

They also found that  $S-X^2$  was most likely to detect misfit when the generating discrimination parameter was highest. This reflects a misfit due to the steepness of the slope. Steepness of slope relates to misfit in the center of the ability distribution, rather than at the extremes, such as that introduced by a non-zero lower asymptote or acceleration parameter. Location of misfit is tied to the distribution of individuals in relation to items.

### **Sensitivity of $S-X^2$ to Prior Informativeness**

Results indicated that the  $S-X^2$  statistic was not sensitive to choice of prior distribution. In the frequentist realm, a study by Sinharay and Lu (2007) indicated that test condition (e.g., average item difficulty, discrimination, and lower asymptote values) did not affect the type I errors or power of the  $S-X^2$  statistic. It could be the case that this statistic is simply less sensitive in computation to misfit, resulting in more uniformly distributed PPP values under the PPC approach.

The fit statistic  $S-X^2$  is computed via Equation (13) with modifications for creating bins that are  $\theta$ -independent, defining bins according to observed test scores (e.g., summated scores) rather than  $\theta$  estimates (Orlando & Thissen, 2000). There are several reasons for this. First,  $\theta$  is a latent variable, preventing the comparison to observed data in a meaningful way (Orlando & Thissen, 2000) except in the case of the 1PL in which the number correct score is a sufficient statistic. Second, grouping examinees into equal-sized groups is highly sample-dependent and the cut-off points, as well as the number of

intervals, affect the resulting statistic. If examinees are grouped according to latent ability level, rather than an equal number of examinees per group, the result is homogenous bins, which also affect the resulting statistic (Orlando & Thissen, 2000).

The  $S-X^2$  statistic avoids this model parameter-dependency, as well as sample dependency, by constructing bins according to number correct scores. Lord and Wingersky (1984) briefly describe the method for predicting joint likelihood distributions for each number correct score. Computing  $S-X^2$  involves the expected proportion of examinees at each score group ( $k$ ) for item  $i$ , calculated via

$$E_{ik1} = \frac{\int P_{i1}(\theta) f^{*i}(k-1|\theta) \phi(\theta) d\theta}{\int f(k|\theta) \phi(\theta) d\theta}, \quad (22)$$

where  $f^{*i}(k-1|\theta)$  and  $f(k|\theta)$  are computed using the recursive algorithm and  $\phi(\theta)$  is the population distribution of  $\theta$  (e.g., the prior). The integrals in Equation 22 are then approximated using Gauss-Hermite quadrature. Computing the expected proportion can be done using only the estimated item parameters for any group of examinees by imposing a distribution of ability parameters, usually Gaussian in nature (Thissen, Pommerich, Billeaud, & Williams, 1995). This eliminates the influence of ability parameter estimation on the model-data fit statistic and may make it less-sensitive overall.

### **Recommendations**

As with other studies of PPC methods, it was found that percent correct does not have an adequate ability to detect model-data misfit, regardless of the type of prior specification used. Informative-inaccurate priors tended to over-classify items as misfitting, resulting in the highest hit rates, but also the highest false positive error rates.

Informative-accurate priors tended to have the highest hit rates, regardless of type of misfit (slope or location on the ability continuum), particularly with *INFIT* as the discrepancy statistic. Noninformative priors (with *INFIT* as the discrepancy statistic) resulted in lower overall hit rates for misfit related to the extremes of the ability continuum. If the location of the misfit is suspected to be in an extreme, then noninformative priors are not recommended.

It will be informative to generate an item-person map, similar to the one produced by Winsteps (Linacre, 2011), mapping summary statistics from ability posterior distributions to summaries of item locations. For items in the general range of the ability parameters, we should have reasonable confidence in the hit rates.

### **Limitations**

One limitation in this study is the use of convergence diagnostics. Diagnostic statistics, such as Geweke, and trace plots, are not a guarantee that the Markov chain has converged. Rather, these methods usually only provide evidence that a chain has not converged (Hoff, 2009). Since Geweke diagnostics were only performed on one replication per condition, the possibility exists that convergence in other conditions was not attained.

Several researchers have recommended that PPC methods be used as pieces of statistical evidence for, rather than a test of, data-model misfit (Berkhof, van Mechelen, & Gelman, 2004; Stern, 2000). This is motivated by recommendations concerning the general practice of using statistical information regarding fit in a larger, theory-guided approach to model criticism (Sinharay, 2005). From this perspective, PPC and PPP

values are viewed as diagnostic measures aimed at assessing model strengths and weaknesses rather than whether the model is true (Fu et al., 2005; Gelman et al., 1996; Levy et al., 2009). I have treated the PPC method as the latter, and I recognize that conclusions would not be made as starkly as “fit or misfit” without more evidence.

Another limitation is in the comparability of this study to other studies on PPC for IRT. Specifically, the estimation methods were different between this study and others (e.g., Sinharay, 2006; Sinharay & Johnson, 2003; Sinharay et al., 2006; and Toribio and Albert, 2011). This study relied upon the Metropolis Hastings algorithm, whereas the previously mentioned studies all used Gibbs sampling. Gibbs sampling is a special case of Metropolis-Hastings where the proposal distribution is constructed differently, the acceptance rate is always one, and Gibbs sampling performs a random walk where at each iteration, the value is randomly updated according to the conditional distribution.

The difference in the algorithms is due to the difference in choice of software and native algorithms to each software program. However, Ames and Samonte (in press) showed the Metropolis-Hastings algorithm used in PROC MCMC can recover item parameters as well as other software programs that use Gibbs sampling, and at a quicker convergence speed.

### **Directions for Future Research**

The use of other discrepancy statistics will be informative to practitioners, particularly those statistics that are sensitive to model misfit in the extremes of the latent ability continuum. The PPC methods did an adequate job of detecting misfit related to slope steepness (i.e., at the center of the latent ability continuum) rather than misfit in the

lower or upper ends of the ability continuum. To control for higher false positive rates, varying the extreme cutoff PPP value should be investigated, as well as how this cutoff might vary with prior specification.

Prior specification was shown to play an important role in the PPC method. In general, noninformative priors had lower hit rates than did informative-accurate prior specifications. Two studies address the effect of priors on the PPC method. In the first, Gelman and colleagues (1996) apply PPC to a study fitting a latent two-class mixture model to the data from an infant temperament study. Dirichlet parameter priors were chosen so that the multinomial probabilities for a variable (e.g., motor activity) were centered on values elicited from psychological theory, but with a large variance. The use of a weak, but not uniform, prior distribution was used to help identify the mixture classes. The authors computed PPP values under a variety of prior distributions. The center of each class of the prior was chosen to either match the values suggested by theory or to represent a uniform distribution. The strength of the prior (informativeness) was also varied.

The authors reported that, as long as the prior distributions were not particularly strong, the center of the prior distribution had little effect on the PPC method and the size of the PPP values remained relatively constant. With incorrect and very strong priors, which are the opposite of noninformative in that they narrow the mass of the likelihood into a smaller region, the PPP value could be quite misleading. However, with correct and very strong priors, the PPP values reflected the true model-data misfit rates of the mixture models. Sinharay and Johnson (2003) commented on the findings of this study,

noting that a strong prior distribution, with reasonable trustworthiness, can be used to more effectively assess the fit of the likelihood part of the model. Despite this recommendation, however, Sinharay and Johnson (2003) used large samples and noninformative priors for their PPC study.

Often, informative priors may prove more useful to IRT practitioners than noninformative priors. For instance, Fox (2010) concluded that the elicited hierarchical prior proved more useful for the 2PL model than did noninformative priors, especially with relatively small datasets when prior information can significantly influence the item parameter estimates. Tsutakawa (1992) used information from a previous years' test administration to help guide the specification of elicited priors. When comparing joint maximum likelihood to Bayesian estimation, Gifford and Swaminathan (1990) found that specification of the priors had only modest effects on the Bayesian estimates, but that the effect of the prior was greater for more complex models, particularly for the lower asymptote parameter of the 3PL model. Similarly, Swaminathan and colleagues (2003) found that the incorporation of ratings provided by subject matter experts produced estimates that were more accurate than those obtained without using such information. The improvement was observed for all item response models, but the improvement was positively related to the number of parameters estimated. Thus, as model complexity grew, the need for specifying informative priors grew in importance (Swaminathan, et al., 2003).

Prior sensitivity analysis is the next stage of research in this area and may provide an alternative to assessing the adequacy of the likelihood model. The basic tool of prior

sensitivity analysis is to change the prior specification and then to recompute the posterior quantity of interest. If there is a practical change in the posterior, the conclusion is that the results are sensitive to the prior specification. If there is no practical change, the data are considered highly informative and the posterior conclusions may be considered indifferent to the prior specification (Albert & Louis, 2000).

Finally, to date, no studies have compared the parameter recovery of IRT model parameters from Metropolis Hastings algorithm to the Gibbs Sampling algorithm. Nor have any studies compared performance of the PPC method for these two algorithms.

## REFERENCES

- Albert, J. A., & Ghosh, M. (2000). Item response modeling. In Dey, D. Ghosh, S., & Mallick, B. (Eds.) *Generalized Linear Models: A Bayesian Perspective*. New York, NY: Marcel Dekker.
- Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics*, *17*, 251-269.
- Ames, A. J. (2014). Posterior predictive checks for polytomous IRT models. Paper presented at the annual conference of the American Educational Researcher Association, April 3-7, Philadelphia, PA.
- Ames, A. J., & Penfield, R. D. (under review). An NCME Instructional Model on Item-Fit Statistics for Item Response Theory Models. *Educational Measurement: Issues and Practice*.
- Ames, A. J., & Samonte, K. M. (in press). Using SAS PROC MCMC for Item Response Theory Models. *Educational and Psychological Measurement*.
- Andrews, M., & Baguley, T. (2013). Prior approval: The growth of Bayesian methods in psychology. *British Journal of Mathematical and Statistical Psychology*, *66*, 1-7.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.
- Baker, F. B. (1987). Methodological review: Item parameter estimation under the one-, two-, and three-parameter logistic models. *Applied Psychological Measurement*, *11*, 111-142.
- Baldwin, P., Bernstein, J., & Wainer, H. (2009). Hip psychometrics. *Statistics in Medicine*, *28*, 2277-2292.
- Bergan, J. R. (2010). Assessing the relative fit of alternative item response theory models to the data. *ATI Research Paper*. Retrieved from [<http://ati-online.com/pdfs/researchK12/AlternativeIRTModels.pdf>].
- Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York, NY: Springer-Verlag.

- Berkhof, J., van Mechelen, I., & Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, *15*, 337-354.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, *Statistical Theories of Mental Test Scores* (chapters 17-20). Reading, MA: Addison-Wesley.
- Birnbaum, A. (1969). Statistical theory for logistic mental test models with a prior distribution of ability. *Journal of Mathematical Psychology*, *6*, 258-276.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.
- Bolfarine, H., & Bazan, J. (2010). Bayesian estimation of the logistic positive exponent IRT model. *Journal of Educational and Behavioral Statistics*, *35*, 693-713.
- Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement*, *51*, 141-162.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 2nd Ed. New York, NY: Routledge.
- Bradlow, E. T., Wainer, H., and Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.
- Carlin, B. P. & Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. London, UK: Chapman and Hall.
- Casella, G., & Berger, R. L. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association*, *82*, 106-111.
- Chib, S. & Greenberg, E. (1995), Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327-335.
- Choo, S., Cohen, A., & Kim, S. (2013). Markov chain Monte Carlo estimation of a mixture item response theory model. *Journal of Statistical Computation and Simulation*, *83*, 278-306.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.
- Curi, M., Singer, J., & Andrade, D. (2011). A model for psychiatric questionnaires with embarrassing items. *Statistical Methods in Medical Research*, *20*, 451-470.
- Curtis, S. M. (2010). BUGS syntax for item response theory. *Journal of Statistical Software*, *36*, 1-34.

- de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York: Guilford Press.
- De Gooijer, J., & Yuan, A. (2011). Some exact tests for manifest properties of latent trait models. *Computational Statistics and Data Analysis*, *55*, 34-44.
- de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115-130.
- de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and interpretation of ancillary variables. *Applied Psychological Measurement*, *33*, 465-485.
- de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267-285.
- de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620-639.
- de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscore. *Applied Psychological Measurement*, *35*, 296-316.
- Drasgow, F., Levine, M., & McLaughlin, M. (1991). Appropriateness measures for some multidimensional test batteries. *Applied Psychological Measurement*, *15*, 171-191.
- Edwards, M. (2010). A Markov chain Monte Carlo approach to confirmatory item factor analysis. *Psychometrika*, *75*, 474-497.
- Efron, B. (1986). Why isn't everyone a Bayesian? *American Statistician*, *40*, 1-5.
- Etnik, R., Fox, J.-P., & van den Hout, A. (2011). A mixture model for the joint analysis of latent development trajectories and survival. *Psychometrika*, *74*, 21-48.
- Etnik, R., Fox, J.-P., & van der Linden, W. J. (2011). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Statistics in Medicine*, *30*, 2310-2325.
- Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*, 893-914.
- Finke, D. (2009). Estimating the effect of nonseparable preferences in Eu treaty negotiations. *Journal of Theoretical Politics*, *21*, 543-569.

- Fox, J. P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. New York, NY: Springer.
- Fraser, C. (1998). NOHARM: A Fortran Program for Fitting Unidimensional and Multidimensional Normal Ogive Models in Latent Trait Theory. Armidale, Australia. The University of New England, Center for Behavioral Studies.
- Fragoso, T., & Curi, M. (2013). Improving psychometric assessment of the Beck Depression Inventory using multidimensional item response theory. *Biometrical Journal*, *55*, 527-540.
- Fu, Z.-H., Tao, J., & Shi, N. (2009). Bayesian estimation of the multidimensional three-parameter logistic model. *Journal of Statistical Computation and Simulation*, *76*, 819-835.
- Geerlings, H., Glas, C., & van der Linden, W. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337-359.
- Gelman, A. (2013). *Running WinBUGS and OpenBUGS from R/S-Plus*. Retrieved from: [<http://cran.r-project.org/web/packages/R2WinBUGS/R2WinBUGS.pdf>].
- Gelman, A., Bois, F., & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, *91*, 1400-1412.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. London, UK: Chapman and Hall.
- Gelman, A., Meng, X.-L., and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-807.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*, 8-38.
- Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317-1339
- Ghosh, M., Ghosh, A., Chen, M.-H., and Agresti, A. (2000). Noninformative priors for one-parameter item response models. *Journal of Statistical Planning and Inference*, *88*, 99-115.
- Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement*, *14*, 33-43.

- Hastings, W. K. (1970), Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.
- Harwell, M., Stone, C., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, 20, 101-125.
- Henson, R. A., Templin, J., Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74, 191-210.
- Hsieh, C.-A., von Eye, A., & Maier, K. (2010). Using a multivariate multilevel polytomous item response model to study parallel processes of change: The dynamic association between adolescents' social isolation and engagement with delinquent peers in the National Youth Survey. *Multivariate Behavioral Research*, 45, 508-552.
- Huang, H., Wang, W., Chen, P., & Su, C. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement*, 37, 619-637.
- Hung, L.-F. (2010). The multigroup multilevel categorical latent growth curve models. *Multivariate Behavioral Research*, 45, 359-392.
- Hung, L.-F. (2011). Formulation and application of the hierarchical generalized random-situation random-weight MIRID. *Multivariate Behavioral Research*, 46, 643-668.
- Hung, L.-F., & Wang, W.-C. (2012). The generalized multilevel facets model for longitudinal data. *Journal of Educational and Behavioral Statistics*, 37, 231-255.
- Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. West Sussex, United Kingdom: John Wiley and Sons, Ltd.
- Jang, E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, 26, 31-73.
- Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49, 82-100.
- Jiao, H., Wang, S., & He, W. (2013). Estimation methods for one-parameter testlet models. *Journal of Educational Measurement*, 50, 186-203.
- Kang, T., & Chen, T. (2008). Performance of the generalized  $S-X^2$  item fit index for polytomous IRT models. *Journal of Educational Measurement*, 45, 391-406.
- Kang, T., Cohen, A., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, 33, 499-518.

- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795.
- Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*, 1343-1370.
- Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*, 399-419.
- Kim, S., Das, S., Chen, M.-H., & Warren, N. (2009). Bayesian structural equation modeling for ordinal response data with missing responses and missing covariates. *Communications in Statistics – Theory and Methods*, *38*, 2748-2768.
- Kim, J.-, S., & Bolt, D. M. (2007). Estimating item response theory models using Markov chain Monte Carlo methods. *Educational Measurement: Issues and Practice*, *26*, 38-51.
- Kloek, T., & van Dijk, H. K. (1978). Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* *46*, 1–19.
- Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*, 2401-2428.
- Levy, R. (2011). Posterior predictive model checking for conjunctive multidimensionality in item response theory. *Journal of Educational and Behavioral Statistics*, *36*, 672-694.
- Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *64*, 208-232.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519–537.
- Li, F., Cohen, A., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353-373.
- Li, Y., & Baser, R. (2012). Using R and WinBUGS to fit a generalized partial credit model for developing and evaluating patient-reported outcomes assessments. *Statistics in Medicine*, *31*, 2010-2016.
- Linacre, J.M. (2011). Winsteps Rasch Measurement Version 3.73 [Software]. Available from: <http://www.winsteps.com>.

- Luo, S., Ma, J., & Kiebertz, K. (2013). Robust Bayesian inference for multivariate longitudinal data by using normal/independent distributions. *Statistics in Medicine*, *32*, 3812-3828.
- Lopes, H. F., & Tobias, J. L. (2011). Confronting prior convictions: On issues of prior sensitivity and likelihood robustness in Bayesian analysis. *Annual Review of Economics*, *3*, 107-131.
- Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). Winbugs- A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325-337.
- Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York, NY: Springer.
- MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, *48*, 188–190.
- Markon, K. (2013). Information utility: Quantifying the total psychometric information provided by a measure. *Psychological Methods*, *18*, 15-35.
- Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, *42*, 1-20.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
- McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49-57.
- Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, *22*, 1142-1160.
- Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*, 335-341.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953), Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.
- Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521-538.

- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177-196.
- Muller, U. K. (2012). Measuring prior sensitivity and prior informativeness in large Bayesian models. *Journal of Monetary Economics*, *59*, 581-597.
- Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.
- Muraki, E., & Bock, D. (1997). PARSCALE: IRT analysis and test scoring for rating scale data [Computer software]. Chicago: Scientific Software.
- Muthén, L. K., & Muthén, B. O. (2011). *MPLUS User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.
- O'Hagan, A. (1994). *Kendall's Advanced Theory of Statistics*. London, UK: Edward Arnold.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*, 50-64.
- Owen, R. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.
- Patz, R., & Junker, B. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.
- R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research). Chicago: The University of Chicago Press.
- Roos, M., & Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *Bayesian Analysis*, *2*, 259-278.
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, *71*, 319-392.
- Saito, M., Iwata, N., Kawakame, N., Matsuyama, Y., & World Mental Health Japan 2002-2003 collaborators (2010). Evaluation of the DSM-IV and ICD-10 criteria

- for depressive disorders in a community population in Japan using item response theory. *International Journal of Methods in Psychiatric Research*, 19, 211-222.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. (*Psychometrika Monograph No. 17*) Richmond, VA: Psychometrics Society.
- Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, 65, 319-335.
- Santos, J., Azevedo, C., & Bolfraine, H. (2013). A multiple group item response theory model with centered skew-normal latent trait distributions under a Bayesian framework. *Journal of Applied Statistics*, 40, 2129-2149.
- SAS Institute Inc. (2014). *The MCMC Procedure*. Cary, NC: SAS Institute Inc.
- Sheng, Y. (2010). A sensitivity analysis of Gibbs sampling for 3PNO IRT models: Effects of prior specifications on parameter estimates. *Behaviormetrika*, 37, 87-110.
- Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, 42, 275-394.
- Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, 59, 429-449.
- Sinharay, S., & Johnson, M. S. (2003). Simulation studies applying posterior predictive model checking for assessing fit of common item response theory models (ETS RR-03-28). Princeton, NJ: Educational Testing Service.
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 298-321.
- Sinharay, S., & Lu, Y. (2007). The correlation between item parameters and item fit statistics. *ETS Research Report RR-07-36*.
- Soares, T., Goncalves, F., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, 34, 348-377.
- Stan Development Team. 2014. Stan: A C++ Library for Probability and Sampling, Version 2.2. Retrieved from: [<http://mc-stan.org>].
- Steinbakk, G. H., & Storbik, G. (2009). Posterior predictive p-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics*, 2, 320-336.

- Stone, C., Ye, F., Zhu, X., & Lane, S. (2009). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*, 63-86.
- Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement, 27*, 27-51.
- Tao, J., Xu, B., Shi, N.-Z., & Jiao, H. (2013). Refining the two-parameter testlet response model by introducing testlet discrimination parameters. *Japanese Psychological Research, 55*, 284-291.
- Thissen, D., Pommerich, M., Billeaud, K., & Williams, V. (1995). Item response theory for scores on tests including polytomous items with ordered responses. *Applied Psychological Measurement, 19*, 39-49.
- Toribio, S. G., & Albert, J. H. (2011). Discrepancy measures for item fit analysis in item response theory. *Journal of Statistical Computation and Simulation, 81*, 1345-1360.
- Tsutakawa, R. K. (1992). Prior distribution for item response curves. *British Journal of Mathematical and Statistical Psychology, 45*, 51-74.
- Tukey, J. W. (1949). Comparing individual means in the analysis of variance. *Biometrics, 5*, 99-114.
- Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39-55.
- Usami, S. (2011). Generalized graded unfolding model with structural equation for subject parameters. *Japanese Psychological Research, 53*, 221-232.
- van den Hout, A., Bockenholt, U., & van der Heijden, P. (2010). Estimating the prevalence of sensitive behaviour and cheating with a dual design for direct questioning and randomized response. *Journal of the Royal Statistical Society, 59*, 723-736.
- Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics, 35*, 5-25.
- Wang, X., Baldwin, S., Wainer, H., Bradlow, E., Reeve, B., Smith, A., Bellizzi, K., & Baumgartner, K. (2010). Using testlet response theory to analyze data from a survey of attitude change among breast cancer survivors. *Statistics in Medicine, 29*, 2018-2044.

- Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26*, 339-352.
- Wright, B., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.
- Yao, L. (2003). BMIRT: Bayesian multivariate item response theory. [Computer software]. Monterey, CA: Defense Manpower Data Center.
- Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*, 339-360.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.
- Zhu, X. (2009). *Checking fit of item response models for performance assessments using Bayesian analysis*. Unpublished dissertation. University of Pittsburgh.
- Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement, 48*, 81-97.