BAYESIAN MODEL CHECKING METHODS FOR DICHOTOMOUS ITEM RESPONSE
THEORY AND TESTLET MODELS

Adam Combs

A Dissertation
Submitted to the Graduate College of Bowling Green State University in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
May 2014

Committee:
Jim Albert, Advisor
Lynne E. Hewitt
Graduate Faculty Representative
Hanfeng Chen
Maria Rizzo
ABSTRACT

Jim Albert, Advisor

The predominant model checking method used in Bayesian item response theory (IRT) models has been the posterior predictive (PP) method. In recent years, two new Bayesian model checking methods have been proposed that may be used as alternatives to the PP method. We refer to these as the prior-predictive posterior simulation (PPPS) method of Dey et al. (1998), and the pivotal discrepancy measure (PDM) method of Johnson (2007). These methods have shown to be effective in other Bayesian models, but have never been implemented with Bayesian IRT models. It is of practical interest to see if either of these two new methods will perform better than the PP method in assessing aspects of fit in an IRT model setting.

In this dissertation, we compared the effectiveness of the PPPS and PDM model checking methods with the PP method in evaluating person fit in two-parameter normal ogive (2PN) IRT models, and overall model goodness-of-fit in 2PN testlet models. Two simulation studies were performed. The first study explored the performance of each method (PP, PPPS, and PDM) in assessing person fit, or the goodness-of-fit of an individual’s set of test answers with the assumed Bayesian 2PN IRT model. Several classical person fit measures were employed under each method. We also introduced using the sum of squared Bayesian latent residuals as a person fit measure. Four different types of person miss-fit were taken from the literature, and response data sets were simulated with certain examinee’s responses following these violations. We found that for most of the measures, the PPPS and PDM methods outperformed the PP method in detecting the examinee’s response patterns simulated to be aberrant under the model. In particular, the sum of squared Bayesian latent residuals
showed to be a very effective measure under the PPPS method.

The second simulation study compares the performance of the PP method and the PPPS method in assessing the overall goodness-of-fit of a Bayesian 2PN IRT model fitted to data generated under a Bayesian 2PN testlet model with equal variance across testlets. Under the PP method we used three goodness-of-fit measures based on biserial correlations that were previously employed for checking the goodness-of-fit of a three-parameter logistic (3PL) IRT model to 3PL testlet data. For use under the PPPS method, we introduced three new goodness-of-fit measures which are calculated from posterior values of the item discrimination parameters. Data sets were simulated under four different values of testlet variance, ranging from very low to fairly high. Looking at the detection rates under the PP method, we saw that the measures performed very poorly in detecting a lack of fit of the 2PN IRT model for all data values of testlet variance. The detection rates of the new measures under the PPPS method showed to be higher than those under the PP method. However, the measures under the PPPS method only showed descent power in detecting lack of fit for large values of data generating testlet variance.
ACKNOWLEDGEMENTS

First and foremost I would like to thank my beautiful wife, Jen, for standing by me and supporting me all the way. I would like to thank her for all the late nights spent watching our children while I worked, and all the Saturday afternoons away which turned into Saturday nights also. Her understanding, love, and encouragement have helped me through. Without her, I would not be here.

I also want to thank Dr. Jim Albert, my advisor, for taking me on as his student. I appreciate Dr. Albert’s guidance in helping me into the field of Bayesian statistics. He was very helpful and patient in helping me select a topic I was interested in. Dr. Albert’s insights and experience have been invaluable in the completion of this dissertation. What’s more was his willingness to share all the “little” things from books to personal experiences that have helped me learn much more than just doing the research alone. Most of all, I want to thank Dr. Albert for his patience and understanding of me trying to complete this work while starting a family. Dr. Albert continued to offer help even after I left BGSU to start a job and finish my dissertation remotely.

I would also like to thank Dr. Hanfeng Chen and Dr. Maria Rizzo. I am not just thanking these two individuals because they served on my dissertation committee, but because they each played an integral role in my journey to the point. While working on my undergraduate degree at BGSU, I had Dr. Chen for several classes. During this time, Dr. Chen took the extra initiative during my visits to his office hours to discuss any possible plans I might have for graduate school. It was Dr. Chen who first inspired me to pursue my master’s degree in mathematics, and at the same time planted the seed of interest in eventually pursuing a doctorate.

When I returned to BGSU to start my doctorate, one of the first classes I took was a
statistical computing class with Dr. Rizzo. I would say that in the course of the work I have
done in this dissertation work, the knowledge I obtained in her class was most valuable. I
also want to thank her for her “R book”, as the graduate students called it, as it has been
a valuable reference for me.

There are two fellow graduate students who I would like to acknowledge. These are Steve
Dinda and Junvie Pailden. When I started my doctorate in statistics I was coming from a
pure mathematics background, and needed to do a lot of catching up. If it was not for the
help of Steve and Junvie, I honestly do not think I would have made it through my first year
of 700-level statistics classes. I want to thank Junvie for all his help in preparing for the
preliminary exams our first year, and for being a wonderful office mate for the remainder of
our time at BGSU. Junvie was always a person I could come to and ask questions, and he
would take the time and patience to answer them. Above all, I would like to thank Steve
and Junvie for being great friends.

I would also like to thank the faculty in the department of Mathematics and Computer
Science at Edinboro University of Pennsylvania. This is where I spent my two years between
finishing my master’s degree and deciding to go back to pursue my doctorate. I made many
great friends while there, Rick White and Frank Marzano to name a few. I want to thank
them all for their encouragement and support initially to go back to graduate school and
pursue the degree, and over the years to finish it. Go Seawolves!

I especially would like to thank my mother and father for offering their understanding
and support all the way through. Lastly, I want to thank my two wonderful children, Gabriel
and Addison. I started this degree for certain reasons, but you are the reason I finished it.
# TABLE OF CONTENTS

## CHAPTER 1  INTRODUCTION  

1.1 Introduction ................................................................................................. 1  
1.2 Introduction to Bayesian Model Checking Methods  
  1.2.1 The Prior Predictive Method ................................................................ 4  
  1.2.2 The Posterior Predictive Method .......................................................... 5  
  1.2.3 New Bayesian Model Checking Methods  
    1.2.3.1 Prior Predictive Posterior Simulation (PPPS) Method ................. 9  
    1.2.3.2 Pivotal Discrepancy Measure (PDM) Method ............................... 12  
1.3 Bayesian Model Checking Procedures Applied to Bayesian Item Response Theory Models  
  1.3.1 Brief Introduction to Item Response Theory Models ......................... 15  
  1.3.2 Model Checking in IRT Models ............................................................ 17  
    1.3.2.1 Research Focus 1: Person Fit Checking in IRT Models ............ 18  
    1.3.2.2 Research Focus 2: Goodness-of-Fit Checking for 2PN Testlet Models ................................................................. 23  

## CHAPTER 2  INTRODUCTION TO ITEM RESPONSE THEORY MODELS  

2.1 Introduction to Item Response Theory ......................................................... 27  
2.2 Item Response Models for Binary Responses  
  2.2.1 One Parameter Model ......................................................................... 29  
  2.2.2 Two-Parameter Model ......................................................................... 30
2.2.3 Three Parameter Model

2.2.4 Model Assumptions

2.3 Item Response Curve

2.4 Introduction to Bayesian Methodology

2.4.1 Introduction

2.4.2 Bayesian IRT Model

2.4.2.1 Likelihood Function

2.4.2.2 Prior Distributions

2.4.2.3 Posterior Distribution of $\theta$, $a$, and $b$

2.4.3 Bayesian Two-Parameter Normal Ogive IRT Model

2.4.3.1 Binary Data Augmentation

2.4.3.2 Bayesian 2PN IRT Model Specification

2.5 Parameter Estimation under Frequentist and Bayesian Approaches

2.5.1 Frequentist Approach

2.5.2 Bayesian Approach

2.5.3 Comparing JMLE and Bayesian Parameter Estimates

2.5.3.1 Similarity of MLEs and Posterior Means

2.5.3.2 Comparison of JML and Bayesian Item and Ability Parameter Estimates

2.5.4 Advantages of the Bayesian Approach

CHAPTER 3 BAYESIAN MODEL CHECKING METHODS

3.1 Introduction

3.2 Bayesian Predictive Model Checking

3.2.1 Prior Predictive Assessment

3.2.2 Posterior Predictive Assessment

3.2.3 Example of Prior Predictive and Posterior Predictive $p$-values

3.2.4 Criticisms of the Prior Predictive and Posterior Predictive Methods
### 3.3 Alternative Bayesian Model Checking Methods: Prior Predictive Posterior Simulation (PPPS) Based Approach

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1 Introduction to the PPPS Approach</td>
<td>74</td>
</tr>
<tr>
<td>3.3.2 Explanation of the PPPS Model Checking Procedure</td>
<td>76</td>
</tr>
<tr>
<td>3.3.3 Example Illustrating the PPPS Model Checking Method</td>
<td>79</td>
</tr>
</tbody>
</table>

### 3.4 Alternative Bayesian Model Checking Methods: Assessment Using Pivotal Discrepancy Measures (PDMs)

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4.1 Introduction to the PDM Method</td>
<td>83</td>
</tr>
<tr>
<td>3.4.2 Background of PDMs</td>
<td>84</td>
</tr>
<tr>
<td>3.4.3 Some Important Theoretical Results for PDMs</td>
<td>85</td>
</tr>
<tr>
<td>3.4.4 Using PDMs for Model Checking</td>
<td>90</td>
</tr>
<tr>
<td>3.4.4.1 Introduction</td>
<td>90</td>
</tr>
<tr>
<td>3.4.4.2 Informal Model Checking Methods using PDMs</td>
<td>92</td>
</tr>
<tr>
<td>3.4.4.3 Formal Model Checking Method using PDMs</td>
<td>94</td>
</tr>
</tbody>
</table>

### 3.5 Applications to IRT Models: A Preview

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5 Applications to IRT Models: A Preview</td>
<td>98</td>
</tr>
</tbody>
</table>

### CHAPTER 4 CHECKING PERSON FIT IN BAYESIAN 2PN IRT MODELS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>100</td>
</tr>
<tr>
<td>4.2 Discussion of Aspects of Fit in IRT Models</td>
<td>101</td>
</tr>
<tr>
<td>4.2.1 Explanation &amp; Background of Person Fit</td>
<td>103</td>
</tr>
<tr>
<td>4.3 Classical Person Fit Measures</td>
<td>104</td>
</tr>
<tr>
<td>4.3.1 Classical Likelihood Based Measures</td>
<td>105</td>
</tr>
<tr>
<td>4.3.1.1 Levine and Rubin (1979)</td>
<td>105</td>
</tr>
<tr>
<td>4.3.1.2 Standardized $L_z$ of Drasgow, Levine &amp; Williams (1985)</td>
<td>105</td>
</tr>
<tr>
<td>4.3.2 Classical Residual Based Measures</td>
<td>106</td>
</tr>
<tr>
<td>4.3.2.1 Unweighted Mean Square Fit Statistic by Wright and Stone (1979)</td>
<td>107</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1.1 Simulation diagram for data sets from the prior predictive distribution under the model ........................................... 5

1.2 Simulation diagram for a posterior predictive data sample $y_1, \ldots, y_N$ .................................................. 7

1.3 Diagram describing the simulation process for the prior predictive posterior simulation (PPPS) method ........................................... 11

1.4 Simulation example illustrating the PPPS method .................................................. 12

1.5 Histograms of posterior samples of pivotal discrepancy measures ........................................... 14

2.1 Example of item response curve under the 2PN IRT model .................................................. 35

2.2 Three item response curves under a 2PN model with the same difficulty of $b = 0$, and different discrimination values .................................................. 36

2.3 Three item response curves with the same discrimination value of $a = 1$, and different difficulty values .................................................. 36

2.4 Three 3PN item response curves with the same discrimination value of $a = 1$, and difficulty value of $b = 0$, but with different guessing parameter values of $c = 0.10, 0.25, \text{ and } 0.50.$ .................................................. 39

2.5 Scatter plots of the 40 true item discrimination parameters (left) and item difficulty parameters (right) and their corresponding joint maximum likelihood estimates .................................................. 50

2.6 Scatter plot of the 200 true ability parameters and their joint maximum likelihood estimates .................................................. 51

2.7 Scatter plots of the 40 true item discrimination parameters (left) and item difficulty parameters (right) and their corresponding posterior estimates .................................................. 55
2.8 Scatter plot of the 200 true ability parameters and their posterior estimates. 55
2.9 Plots showing the prior density, likelihood, and posterior density, with increasing prior variance 57
2.10 Plots showing the prior density, likelihood, and posterior density, with increasing sample size 57
2.11 Difference plot for item discrimination parameter estimates 59
2.12 Difference plot for item difficulty parameter estimates 60
2.13 Difference plot for item ability parameter estimates 61
3.1 Histograms of the prior predictive and posterior predictive samples of the discrepancy measure $T(y) = \sum_1^n y_i/n$ simulated under the model $y_i \sim N(\theta, 1)$; $\theta \sim N(0, 1)$ 70
3.2 Histograms of the prior predictive and posterior predictive p-values computed from the samples of the discrepancy measure $T(y) = \sum_1^n y_i/n$ 72
3.3 Comparison of posterior sample densities generated from prior predictive data under the model to the posterior sample density of the observed data 81
3.4 Normal quantile plots of 1,000 simulated values of $d_1(y_i, \hat{\theta}_i) = (y_i - \hat{\theta}_i)/\sigma$ for each $i = 1, \ldots, 10$ 88
3.5 Tukey mean-difference plot for the quantiles plotted in figure 3.4 89
3.6 Normal quantile plots of 1,000 simulated values of $d_2(i, \tilde{\mu}) = (\tilde{\theta}_i - \tilde{\mu})/\tau$ for each $i = 1, \ldots, 10$ 89
3.7 Tukey mean-difference plot for the quantiles plotted in figure 3.6 90
3.8 Histograms of posterior samples of discrepancy measure $d_1(y, \theta) = \sum_{i=1}^{10} (y_i - \theta_i)^2/\sigma^2$ with data generated under the true model and false model 93
4.1 Density and 90% probability interval plots for Bayesian residuals 131
4.2 Density and 90% probability interval plots for Bayesian latent residuals 132
4.3 Posterior samples of person fit measure $d_Z$ for examinee 3 (non-guesser) and examinee 10 (guesser) are graphed in histograms. ................................. 135

4.4 Histograms of 500 sample values of the measure $t(UB_i)$, for three randomly selected examinees $i = 39, 85, \text{and} 147$, simulated under the 1PN IRT model ........................................... 139

4.5 90% probability interval plots of Bayesian latent residuals for 6 examinees identified with misfitting response patterns by $L_z$ ................................................. 140

4.6 Samples of 1,000 values of measure $L_z$ computed using simulated data and parameters under the model .......................................................... 148

4.7 A dot plot for each response violation type shows the average detection rates of each person fit measure calculated while performing the posterior predictive model checking method. .............................................. 169

4.8 A dot plot for each response violation type shows the mean detection rates .......................................................... 172

4.9 A dot plot for each response violation type shows the detection rates of each of the person fit measures calculated while performing the prior predictive posterior simulation (PPPS) model checking method. .............................................. 174

4.10 Parallel bar charts for the Guessing and Sleeping response violations ........ 180

4.11 Parallel bar charts for the Random Response and Cheating response violations 183

5.1 Difference plots for the posterior item discrimination parameter estimates computed from fitting the 2PN testlet model and 2PN IRT model to data generated from a 2PN testlet model .............................................. 196

5.2 Difference plots for the posterior item difficulty parameter estimates computed from fitting the 2PN testlet model and 2PN IRT model to data generated from a 2PN testlet model .............................................. 197

5.3 Difference plots for the posterior ability parameter estimates computed from fitting the 2PN testlet model and 2PN IRT model to data generated from a 2PN testlet model .............................................. 199
5.4 Histogram of a posterior predictive sample of the standard deviations of sample point biserial correlations. .................................................. 208

5.5 Scatter plots of true values against posterior estimates of the item discrimination parameters $a_k$ ................................................................. 212

5.6 Histograms of posterior predictive samples of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$ .............................................. 214

5.7 Box plots giving the 90% posterior intervals of the samples of the computed standard deviation on the discrimination parameters, denoted as $sd(a)$ . . . . 219

5.8 Plot of the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of Figure 5.7) of each posterior sample of the measure $sd(a)$ ................................................................. 220

5.9 Box plots giving the 90% posterior intervals of the samples of the computed maximum on the discrimination parameters, denoted as $max(a)$ . . . . . . . 221

5.10 Plot of the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of Figure 5.9) of each posterior sample of the measure $max(a)$ ................................................................. 222

5.11 Box plots giving the 90% posterior intervals of the samples of the computed mean on the discrimination parameters, denoted as $mean(a)$ . . . . . . . 223

5.12 Plot of the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of Figure 5.11) of each posterior sample of the measure $mean(a)$ ................................................................. 224
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Computed correlation coefficients between the set of true discrimination</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>parameters ( a ), difficulty parameters ( b ), and ability parameters ( \theta ), and their respective sets of joint maximum likelihood and Bayesian estimates.</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Empirical ( p )-values computed using the discrepancy measure ( d_{i,i}(y_i, \theta_i) = y_i - \theta_i ), ( i = 1, \ldots, 10 ), to check for outlying values</td>
<td>82</td>
</tr>
<tr>
<td>4.1</td>
<td>Examples of different response patterns.</td>
<td>103</td>
</tr>
<tr>
<td>4.2</td>
<td>Simulated probabilities of a correct response to 5 items for a given examinee.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>The variance ( v_{ik} = p_{ik}(1 - p_{ik}) ) is also computed for each item.</td>
<td>110</td>
</tr>
<tr>
<td>4.3</td>
<td>Item number and JMLE's for the item discrimination and difficulty parameters, response pattern for examinee 50 and estimated probabilities of success</td>
<td>119</td>
</tr>
<tr>
<td>4.4</td>
<td>Ranked difficulty parameter mle's</td>
<td>123</td>
</tr>
<tr>
<td>4.5</td>
<td>Column means ( G_k, k = 1, \ldots, 20 ), of the estimated probability matrix ( \hat{P} )</td>
<td>124</td>
</tr>
<tr>
<td>4.6</td>
<td>Average type I error rates over the 200 examinees for each person fit measure evaluated under each method: PP, PPPS, and PDM</td>
<td>165</td>
</tr>
<tr>
<td>4.7</td>
<td>Mean detection rates of the person fit measures evaluated using the posterior predictive method</td>
<td>168</td>
</tr>
<tr>
<td>4.8</td>
<td>Mean detection rates of the person fit measures ( d_Z ) and ( L_z ) evaluated using the pivotal discrepancy measure (PDM) method</td>
<td>171</td>
</tr>
<tr>
<td>4.9</td>
<td>Mean detection rates of the person fit measures evaluated using the prior posterior predictive simulation (PPPS) method</td>
<td>173</td>
</tr>
<tr>
<td>4.10</td>
<td>Average rates of the person fit measures detecting a Guessing response viola-</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>tion over the 10 violating examinees</td>
<td></td>
</tr>
</tbody>
</table>
4.11 Average rates of the person fit measures detecting Sleeping over the 10 violating examinees .......................................................... 179
4.12 Average rates of the person fit measures detecting Random Responding over the 10 violating examinees ..................................................... 180
4.13 Average rates of the person fit measures detecting Cheating over the 10 violating examinees .............................................................. 182

5.1 Mean absolute errors (i.e. MAE = mean|estimate - true value|) computed between the true parameter values and posterior estimates of the examinee ability and item parameters ............................................. 198
5.2 Mean Posterior Interval Width (MPIW = \(\sum_{i=1}^{500}(\hat{\theta}_{i,0.975} - \hat{\theta}_{i,0.025})/500\)) of the sample 95% posterior intervals for the examinee abilities .............................................. 201
5.3 Percentage of true values \(\theta_i\) that were contained in the sample 95% posterior intervals computed from fitting the 2PN IRT model .......................................................... 201
5.4 Type I error rates, or number of times out of 100 simulations the 3 discrepancy measures \(sd(\hat{r}_{bis}), max(\hat{r}_{bis}),\) and \(mean(\hat{r}_{bis})\) ................................................................. 227
5.5 Detection rates for simulation study using the PP method .............................................................. 228
5.6 Type I Error Rates for simulation study under PPPS method .......................................................... 230
5.7 Detection Rates for simulation study under PPPS method ............................................................. 231
CHAPTER 1
INTRODUCTION

1.1 Introduction

The purpose of this dissertation is to study model checking procedures in Bayesian item response theory (IRT) models. So far, the predominant Bayesian model checking method has been the posterior predictive method. There have been two newer Bayesian model checking methods that have arisen in recent years, and we wish to implement these to check aspects of fit with IRT models.

In this chapter we will introduce our research focus and specific research questions. This chapter will contain a cursory introductory to material on Bayesian model checking methods and IRT models. These topics will be expounded on in much more detail in later chapters. The quick overview of the topics in this section is meant to give the reader just enough knowledge to understand the research questions we will present.

The rest of the dissertation will be outlined as follows. Chapter 2 will give an introduction to standard IRT models. Chapter 3 will discuss some current model checking procedures that are being employed in Bayesian model checking, and then introduce two new methods. Chapters 4 and 5 will discuss the main results of the research that was conducted by implementing the new model checking methods. The work done in these chapters will look at applying the Bayesian model checking procedures discussed in chapter 3 to Bayesian IRT models. Specifically, chapter 4 will investigate how the different procedures perform in assessing person fit in two-parameter normal ogive (2PN) Bayesian IRT models. Chapter 5 will apply some of the model checking procedures towards assessing model fit of a Bayesian 2PN IRT model to data that was generated under Bayesian 2PN testlet model.
1.2 Introduction to Bayesian Model Checking

Methods

In statistical modeling it is fundamental to assess how well a proposed model fits a set of observed data. This may be done as a preliminary step in an overall model selection process. In this, several models may be posited and before undergoing procedures to choose a best fitting model, we want to see if any of the models should be excluded because they do not fit the data well. In this type of model checking we are concerned with the overall goodness-of-fit of a model to a set of data. When performing model checking, or model diagnostics, we can also check whether any assumptions the model makes about the data fail to hold.

In classical model checking we begin by choosing some test statistic $T(Y)$ that captures the aspect of fit we want to check (e.g. overall goodness-of-fit). Depending on the aspect of fit being checked, large or small values of the statistic $T$ evaluated at the observed data $y_{obs}$ indicate misfit. For example, suppose we are checking the goodness-of-fit of a proposed sampling density $f(y|\theta)$ for $y_{obs}$, i.e. we are testing the hypothesis $H_0 : y_{obs} \sim f(y|\theta)$. We assess how “large” or “small” the value of $T(y_{obs})$ is by seeing where it lies within its sampling distribution derived under the proposed model. This is typically done by computing a tail area probability, or $p$-value given by,

$$p(y_{obs}) = P(T(Y) \geq T(y_{obs})|H_0).$$  

(1.1)

Values of $p(y_{obs})$ significantly close to zero or one indicate that the value of $T(y_{obs})$ is unlikely to be observed if the model were true. This is evidence against $y_{obs}$ having been generated under $f(y|\theta)$, or that $f(y|\theta)$ is not a good fitting model for $y_{obs}$.

Next we will discuss some Bayesian model checking methods for Bayesian models. Before we do this however, we will give a quick run through of what a basic Bayesian model looks like. This is meant to give the reader just enough understanding to follow the proceeding
discussions. A more detailed discussion of Bayesian methodology will be given in chapter 3. Under the Bayesian paradigm, parameters may be viewed as random quantities and may be thought to have probability distributions just like data. A Bayesian model has two basic components, a prior distribution for the parameters and a sampling density for the data. For the parameter $\theta$, the prior distribution $p(\theta)$ is the distribution that $\theta$ is believed to follow prior to observing any data. In this, $p(\theta)$ represents our prior beliefs and knowledge we have about the parameter $\theta$ before we observe $y$. The data $y$ has a sampling density $f(y|\theta)$, which as a function of $\theta$ is the likelihood function. So, a Bayesian model would be specified by giving the sampling density for the data, and prior distributions of any parameters. For example,

$$y \sim f(y|\theta) \quad (1.2)$$
$$\theta \sim p(\theta), \quad (1.3)$$

would represent a Bayesian model specification.

A very important aspect of a Bayesian model is the posterior distribution that comes from it. Once the data is observed we have more knowledge about the parameter $\theta$ than when we did prior to observing it, or when we just had $p(\theta)$. At this point a posterior distribution of $\theta$ is derived, denoted $\pi(\theta|y)$. The posterior distribution is obtained using Bayes’ Theorem, from which we have that

$$\pi(\theta|y) \propto f(y|\theta)p(\theta), \quad (1.4)$$

where the posterior distribution is proportional to the likelihood times the prior density up to a normalizing constant (see chapter 2 section 2.4 for full details). This distribution is important because, given the Bayesian model, this will be the distribution from which inferences about $\theta$ will be made.

We now discuss model checking in Bayesian models. Under the classical approach, the
main idea is that if \( y_{\text{obs}} \) was generated under the model then the observed values of the test statistic should be likely or typical under its sampling distribution. A concept in Bayesian model checking is that if a model fits a set of data \( y_{\text{obs}} \), then replicated data generated under the model should be similar to \( y_{\text{obs}} \) (Gelman et al. 2004). Two methods based on this are prior predictive and posterior predictive model checking.

1.2.1 The Prior Predictive Method

The prior predictive method checks the similarity between the observed data and data simulated from the prior predictive distribution under the model (Box 1980). The prior predictive distribution of \( y \), also marginal distribution, is defined by the density

\[
g(y) = \int f(y|\theta)p(\theta)d\theta, \tag{1.5}
\]

where \( f(y|\theta) \) is the sampling density of \( y \) and \( p(\theta) \) is the prior distribution of \( \theta \). To “compare” the observed data \( y_{\text{obs}} \) to \( g(y) \) we also use test quantities \( T \). We compare the observed value \( T(y_{\text{obs}}) \) to the prior predictive distribution of \( T \). As in the classical setting, a tail area probability or prior predictive \( p \)-value is computed,

\[
p_{\text{prior}}(y_{\text{obs}}) = P(T(y) \geq T(y_{\text{obs}})). \tag{1.6}
\]

This \( p \)-value can be estimated by simulating many data sets \( y_r, r = 1, \ldots, R \), from \( g(y) \) and computing \( T(y_1), \ldots, T(y_R) \). Then we look at the proportion of values \( T(y_r), r = 1, \ldots, R \), greater than \( T(y_{\text{obs}}) \). A prior predictive data set \( y \) can be simulated from \( g(y) \) by (1) simulating parameter(s) \( \theta \) from its prior distribution \( p(\theta) \) and (2) simulating \( y \) from its sampling density \( f(y|\theta) \), given the prior value of \( \theta \). This process is illustrated in figure 1.1.
Prior Predictive Data Simulation

\[
p(\theta) \xrightarrow{\theta_1} f(y|\theta_1) \rightarrow y_1 \quad \text{Prior Predictive Data Samples} \\
\quad \theta_2 \xrightarrow{f(y|\theta_2)} y_2 \\
\quad \vdots \\
\quad \theta_R \xrightarrow{f(y|\theta_R)} y_R
\]

Figure 1.1 Simulation diagram for data sets from the prior predictive distribution under the model. First, a set of parameters \( \theta_1, \ldots, \theta_R \) is generated from the prior distribution \( p(\theta) \). Second, for \( r = 1, \ldots, R \), we simulate a data sample \( y_r \) from the sampling density \( f(y|\theta_r) \) given \( \theta_r \).

1.2.2 The Posterior Predictive Method

The posterior predictive method checks model fit by looking at the similarity between the observed data and data drawn from the posterior predictive distribution under the model (Gelman et al. 1996). The posterior predictive distribution of \( y \), given \( y_{\text{obs}} \), is defined as,

\[
m(\tilde{y}|y_{\text{obs}}) = \int f(\tilde{y}|\theta)\pi(\theta|y_{\text{obs}})d\theta,
\]

where \( m(\tilde{y}|y_{\text{obs}}) \) is the distribution of posterior predictive data \( \tilde{y} \) given \( y_{\text{obs}} \), \( f(\tilde{y}|\theta) \) is the sampling density of \( \tilde{y} \), and \( \pi(\theta|y_{\text{obs}}) \) is the posterior distribution of \( \theta \) given \( y_{\text{obs}} \).

As in the prior predictive method, we use some test quantity \( T \) and compare \( T(y_{\text{obs}}) \) with the posterior predictive distribution of \( T \). This is done by computing the tail area probability, or posterior predictive \( p \)-value, as

\[
p_{\text{post}}(y_{\text{obs}}) = P(T(\tilde{y}) \geq T(y_{\text{obs}})).
\]

This can be estimated by simulating posterior predictive data sets \( \tilde{y}_r \sim m(\tilde{y}|y_{\text{obs}}) \), \( r = 1, \ldots, R \), and computing \( T(\tilde{y}_r) \). We then look at the proportion of values \( T(\tilde{y}_r) \), \( r = 1, \ldots, R \),
that are greater than then $T(y_{obs})$. Posterior predictive data values can be simulated by (1) simulating $\hat{\theta}$ from the posterior distribution $\pi(\theta|y_{obs})$ and (2) generating data $\tilde{y}$ from the sampling density $f(y|\hat{\theta})$ given the posterior value $\hat{\theta}$.

In the posterior predictive method, it is useful to define measures that are functions of the data and parameters (Gelman et al. 1996). In this, we would have discrepancy measures $T(y, \theta)$ that depend on both $y$ and $\theta$. In using these we compute a posterior sample of $T(y_{obs}, \hat{\theta})$ by simulating posterior parameters $\hat{\theta}_1, \ldots, \hat{\theta}_N$ from the posterior distribution $\pi(\theta|y_{obs})$ and computing $T(y_{obs}, \hat{\theta}_1), \ldots, T(y_{obs}, \hat{\theta}_N)$. This sample is then compared to the joint distribution of $\tilde{y}$ and $\hat{\theta}$. Similar to (1.8), we do this by computing a tail area probability,

$$p_{post}(y_{obs}, \hat{\theta}) = P[T(\tilde{y}, \hat{\theta}) \geq T(y_{obs}, \hat{\theta})|y_{obs}].$$

(1.9)

Given our posterior parameter sample $\hat{\theta}_1, \ldots, \hat{\theta}_N$, we estimate the $p$-value in (1.9) by simulating posterior predictive data sets $\tilde{y}_1, \ldots, \tilde{y}_N$ from the sampling density $f(y|\hat{\theta}_n)$, $n = 1, \ldots, N$, respectively. We then compute $T(\tilde{y}_1, \hat{\theta}_1), \ldots, T(\tilde{y}_N, \hat{\theta}_N)$ and look at the proportion of values $T(\tilde{y}_n, \hat{\theta}_n)$ greater than $T(y_{obs}, \hat{\theta}_n)$ for $n = 1, \ldots, N$. We illustrate this in figure 1.2.

For reasons that will be discussed in chapter 3, many Bayesian researchers prefer using the posterior predictive method over the prior predictive method when doing model checking. One brief reason is that the posterior predictive method is easy to implement with a wide range of models, where the prior predictive method can be limited to the types of models it can be used for. As will also be discussed, some Bayesians do have criticisms of the posterior predictive method. Among a few are that, depending on the model, it can tend to be a conservative testing procedure and its $p$-values can be difficult to interpret (Bayarri and Berger 1999, Bayarri and Berger 2000).

We give a brief example to illustrate the use of the posterior predictive method. Suppose
Posterior Predictive Simulation

\[ y_{\text{obs}} \rightarrow \pi(\theta|y_{\text{obs}}) \]

\[ \tilde{\theta}_1 \xrightarrow{\text{f}} \tilde{y}_1 \xrightarrow{\text{T}} \]

\[ \tilde{\theta}_2 \xrightarrow{\text{f}} \tilde{y}_2 \xrightarrow{\text{T}} \]

\[ \vdots \]

\[ \tilde{\theta}_N \xrightarrow{\text{f}} \tilde{y}_N \xrightarrow{\text{T}} \]

\[ \text{Posterior Predictive Sample of } T \]

**Estimation of Posterior Predictive p-value**

\[ p_{\text{post}}(y_{\text{obs}}, \tilde{\theta}) \approx \frac{\sum_{n=1}^{N} I[T(\tilde{y}_n, \tilde{\theta}_n) \geq T(y_{\text{obs}}, \tilde{\theta}_n)]}{N} \]

Figure 1.2 Simulation diagram for a posterior predictive data sample \( \tilde{y}_1, \ldots, \tilde{y}_N \) and posterior predictive sample of the discrepancy measure \( T(\tilde{y}_1, \tilde{\theta}_1), \ldots, T(\tilde{y}_N, \tilde{\theta}_N) \). First, we simulate a posterior parameter sample \( \tilde{\theta}_1, \ldots, \tilde{\theta}_N \) from \( \pi(\theta|y_{\text{obs}}) \), and then we simulate posterior predictive data sets \( \tilde{y}_n \) from the sampling density \( f(y|\tilde{\theta}_n) \), for \( n = 1, \ldots, N \). We then compute \( T(\tilde{y}_1, \tilde{\theta}_1), \ldots, T(\tilde{y}_N, \tilde{\theta}_N) \), and estimate the \( p \)-value in (1.9) by looking at the proportion of values \( T(\tilde{y}_n, \tilde{\theta}_n) \) greater than \( T(y_{\text{obs}}, \tilde{\theta}_n) \) for \( n = 1, \ldots, N \).

we have the Bayesian model,

\[ y_i \sim N(\theta, 1), \quad i = 1, \ldots, n \]  
\[ \theta \sim N(0, 1). \]  

(1.10)

Given observed data \( y_{\text{obs}} \), we want to check the goodness-of-fit of the model in (1.10). We will use the test quantity \( T(y, \theta) = \bar{y} - \theta \). This is an example of a discrepancy measure that is a function of both data and parameters. We should note here that all test quantities, or statistics, in a classical setting are functions of the data only. In a Bayesian setting we can use measures like \( T(y, \theta) \) because, as mentioned, we can simulate posterior values \( \tilde{\theta} \) from \( \pi(\theta|y_{\text{obs}}) \) and actually compute a value of \( T \) using these, i.e. \( T(y_{\text{obs}}, \tilde{\theta}) = \bar{y}_{\text{obs}} - \tilde{\theta} \).

The posterior distribution of \( \theta, \pi(\theta|y) \), corresponding to the model in (1.10) can be found
in closed form. In general though, we can simulate parameter values from the posterior distribution using Markov chain Monte Carlo (MCMC) methods. To be concise here we will not describe the simulation details, but focus on the main points of the model checking process.

For our example we simulate one observed data set \( y^{(\text{true})}_{\text{obs}} \) under the model in (1.10), and another \( y^{(\text{false})}_{\text{obs}} \) under the model: \( y_i \sim N(\theta, 1); \theta \sim N(3, 1) \). As described above, given \( y_{\text{obs}} \) we simulate \( N = 1,000 \) posterior parameter values \( \tilde{\theta}_1, \ldots, \tilde{\theta}_N \sim \pi(\theta | y_{\text{obs}}) \) and compute the posterior sample \( T(y_{\text{obs}}, \tilde{\theta}_1), \ldots, T(y_{\text{obs}}, \tilde{\theta}_N) \), where \( T(y_{\text{obs}}, \tilde{\theta}_n) = \bar{y}_{\text{obs}} - \tilde{\theta}_n \). We then simulate posterior predictive data \( \tilde{y}_n \sim f(y | \tilde{\theta}_n) = N(\tilde{\theta}_n, 1), n = 1, \ldots, N \), and compute \( T(\tilde{y}_1, \tilde{\theta}_1), \ldots, T(\tilde{y}_N, \tilde{\theta}_N) \), where \( T(\tilde{y}_n, \tilde{\theta}_n) = \bar{\tilde{y}}_n - \tilde{\theta}_n \). These computations are done for both \( y^{(\text{true})}_{\text{obs}} \) and \( y^{(\text{false})}_{\text{obs}} \). For the data set \( y^{(\text{true})}_{\text{obs}} \), we compute,

\[
p_{\text{post}}(y^{(\text{true})}_{\text{obs}}, \tilde{\theta}) \approx \frac{\sum_{n=1}^{N} I[T(\bar{y}_n, \tilde{\theta}_n) \geq T(y^{(\text{true})}_{\text{obs}}, \tilde{\theta}_n)]}{N} = 0.591.
\]  

For the data set \( y^{(\text{false})}_{\text{obs}} \), we compute,

\[
p_{\text{post}}(y^{(\text{false})}_{\text{obs}}, \tilde{\theta}) \approx \frac{\sum_{n=1}^{N} I[T(\bar{y}_n, \tilde{\theta}_n) \geq T(y^{(\text{false})}_{\text{obs}}, \tilde{\theta}_n)]}{N} = 0.026.
\]  

Looking at (1.11), this indicates that the values of \( T(y^{(\text{true})}_{\text{obs}}, \tilde{\theta}_n) \) were less than the corresponding values of \( T(\bar{y}_n, \tilde{\theta}_n) \) about 59% of the time. In (1.12) we see that the values of \( T(y^{(\text{false})}_{\text{obs}}, \tilde{\theta}_n) \) were less than the corresponding values of \( T(\bar{y}_n, \tilde{\theta}_n) \) only about 2.6% of the time. This means that overall most of the sample values of \( T(y^{(\text{false})}_{\text{obs}}, \tilde{\theta}_n) \) were larger than \( T(\bar{y}_n, \tilde{\theta}_n) \), which indicates a discrepancy between the sample \( T(y^{(\text{false})}_{\text{obs}}, \tilde{\theta}_n), n = 1, \ldots, N \) and the sample \( T(\bar{y}_n, \tilde{\theta}_n), n = 1, \ldots, N \), computed from the joint distribution of \((\tilde{y}, \tilde{\theta})\). This in turn offers evidence that the data \( y^{(\text{false})}_{\text{obs}} \) was not generated under the model.
1.2.3 New Bayesian Model Checking Methods

1.2.3.1 Prior Predictive Posterior Simulation (PPPS) Method

There have been two alternative Bayesian model checking methods proposed. The first is what we will call the prior predictive posterior simulation (PPPS) method. This method was proposed by [Dey et al. (1998)](cite:Dey1998). Recall that the main idea behind the prior predictive and posterior predictive methods is that if the observed data $y_{obs}$ was generated under the model then it should look “similar” to data generated under the prior predictive or posterior predictive distribution, respectively. The approach by [Dey et al. (1998)](cite:Dey1998) is similar in concept. The main idea that [Dey et al. (1998)](cite:Dey1998) proposes is that if $y_{obs}$ was produced by the model then the posterior distribution $\pi(\theta|y_{obs})$ we get given $y_{obs}$ should be similar to the posterior distributions we would get from data simulated under the model.

Specifically, we simulate data sets $y_1, \ldots, y_R$ from the prior predictive distribution under the model and compute the corresponding posterior distributions $\pi(\theta|y_1), \ldots, \pi(\theta|y_R)$. Then, if $y_{obs}$ was generated under the model, $\pi(\theta|y_{obs})$ should be similar to these other posterior distributions. In practice we still use test quantities $T$ since these can be used to capture different aspects of model fit. We then compare the posterior distribution of $T$ given $y_{obs}$, to the posterior distributions of $T$ given $y_r$, $r = 1, \ldots, R$.

We give an example using the same data sets $y_{obs}^{(true)}$ and $y_{obs}^{(false)}$ generated earlier, and the same discrepancy measure $T(y, \theta) = \bar{y} - \theta$. In the PPPS method, we would start as we did in the previous example. In the previous example we simulated posterior values denoted now as $\hat{\theta}_1^{(0)}, \ldots, \hat{\theta}_N^{(0)} \sim \pi(\theta|y_{obs})$, from the posterior distribution of $\theta$ given $y_{obs}^{(true)}$ and $y_{obs}^{(false)}$, respectively. We then computed a sample $T(y_{obs}, \hat{\theta}_1^{(0)}), \ldots, T(y_{obs}, \hat{\theta}_N^{(0)})$ from the posterior distribution of $T$ given $y_{obs}$, $y_{obs}^{(true)}$ and $y_{obs}^{(false)}$, respectively.

Now, we can simulate $y_1, \ldots, y_R$ from the prior predictive distribution under the true model, and for each $r = 1, \ldots, R$ we can draw a posterior sample $\tilde{\theta}_1^{(r)}, \ldots, \tilde{\theta}_N^{(r)} \sim \pi(\theta|y_r)$ from the posterior distribution of $\theta$ given $y_r$. We then compute the posterior sample of $T$: 
\( T(y_r, \tilde{\theta}_1^{(r)}), \ldots, T(y_r, \tilde{\theta}_N^{(r)}) \). Once we get the \( R \) different posterior samples of \( T \) given \( y_r \), we compare these to the sample based on \( y_{\text{obs}} \). We illustrate this overall simulation process of the PPPS method in figure 1.3.

For our example we simulate \( R = 50 \) prior predictive data sets \( y_1, \ldots, y_{50} \) under the (true) model in (1.10). With these we compute the corresponding 50 posterior samples of \( T \) given \( y_r, r = 1, \ldots, 50 \). As mentioned, in the PPPS method we want to compare the posterior sample of \( T \) given the observed data to the \( R \) posterior samples of \( T \) given the simulated data \( y_r \) under the model. For now, we will just compare the samples graphically. We will discuss a more precise method for quantitatively comparing the samples in chapter 3.

In the left and right sides of figure 1.4 we plot the sample densities of all the posterior samples of \( T \) given \( y_r, r = 1, \ldots, 50 \), in light color. The posterior sample densities of \( T \) given \( y_{\text{obs}}^{(\text{true})} \) and \( y_{\text{obs}}^{(\text{false})} \) are plotted in dark color in the left and right sides of figure 1.4 respectively. In the left side, we see that the posterior sample of \( T \) given \( y_{\text{obs}}^{(\text{true})} \) is very similar to the posterior samples produced from the data \( y_1, \ldots, y_{50} \) replicated under the model. In the right side, we see the sample density of \( T \) given \( y_{\text{obs}}^{(\text{false})} \) is not very similar, being shifted to the right of the other sample densities. Since the posterior sample of the measure \( T \) given \( y_{\text{obs}}^{(\text{false})} \) is atypical in reference to the posterior samples of \( T \) given the data generated under the model, this indicates that \( y_{\text{obs}}^{(\text{false})} \) was not generated under the model in (1.10).

Some of the noted advantages of this method over the posterior predictive method is that it may be less conservative and possibly more powerful. A drawback of the PPPS method is that, depending on the model, the simulation time may be high. This comes from the fact that after simulating the \( R \) prior predictive data sets \( y_1, \ldots, y_R \) under the model, we are then simulating samples of posterior parameters from each posterior distribution given \( y_r \). If the posterior distributions are known in closed form, this will not add much cost in simulation time. If they are not known in closed form, and we are using MCMC procedures to simulate the samples, this may add to the overall time cost.
Posterior Sample Given $y_{obs}$

$y_{obs} \xrightarrow{} \pi(\theta | y_{obs})$

\[
\begin{align*}
\tilde{\theta}_1^0 & \xrightarrow{} T(y_{obs}, \tilde{\theta}_1^0) \\
\vdots & \vdots \\
\tilde{\theta}_N^0 & \xrightarrow{} T(y_{obs}, \tilde{\theta}_N^0)
\end{align*}
\]

Posterior Sample given $y_{obs}$

Prior Predictive Posterior Simulation

$y_1, \ldots, y_R \sim g(y) = \int f(y|\theta)p(\theta)d\theta$ : Draw $R$ Prior Predictive Data Sets

$y_1 \xrightarrow{} \pi(\theta | y_1)$

\[
\begin{align*}
\tilde{\theta}_1^{(1)} & \xrightarrow{} T(y_1, \tilde{\theta}_1^{(1)}) \\
\tilde{\theta}_2^{(1)} & \xrightarrow{} T(y_1, \tilde{\theta}_2^{(1)}) \\
\vdots & \vdots \\
\tilde{\theta}_N^{(1)} & \xrightarrow{} T(y_1, \tilde{\theta}_N^{(1)})
\end{align*}
\]

PPP Sample 1

$y_2 \xrightarrow{} \pi(\theta | y_2)$

\[
\begin{align*}
\tilde{\theta}_1^{(2)} & \xrightarrow{} T(y_2, \tilde{\theta}_1^{(2)}) \\
\tilde{\theta}_2^{(2)} & \xrightarrow{} T(y_2, \tilde{\theta}_2^{(2)}) \\
\vdots & \vdots \\
\tilde{\theta}_N^{(2)} & \xrightarrow{} T(y_2, \tilde{\theta}_N^{(2)})
\end{align*}
\]

PPP Sample 2

$\vdots$

$\vdots$

$\vdots$

$y_R \xrightarrow{} \pi(\theta | y_R)$

\[
\begin{align*}
\tilde{\theta}_1^{(R)} & \xrightarrow{} T(y_R, \tilde{\theta}_1^{(R)}) \\
\tilde{\theta}_2^{(R)} & \xrightarrow{} T(y_R, \tilde{\theta}_2^{(R)}) \\
\vdots & \vdots \\
\tilde{\theta}_N^{(R)} & \xrightarrow{} T(y_R, \tilde{\theta}_N^{(R)})
\end{align*}
\]

PPP Sample $R$

Figure 1.3 Diagram describing the simulation process for the prior predictive posterior simulation (PPPS) method. (Top) We compute a posterior sample of $T(\mathbf{y}, \theta)$ from simulated posterior parameters $\tilde{\theta}_1^0, \ldots, \tilde{\theta}_N^0 \sim \pi(\theta | y_{obs})$, given $y_{obs}$. (Bottom) We simulate prior predictive data sets $\mathbf{y}_1, \ldots, \mathbf{y}_R$ and compute posterior samples of $T$ from simulated posterior parameters $\tilde{\theta}_1^r, \ldots, \tilde{\theta}_N^r \sim \pi(\theta | \mathbf{y}_r)$ given $\mathbf{y}_r$, $r = 1, \ldots, R$. 
Figure 1.4 After simulating 50 prior predictive data sets $y_1, \ldots, y_{50}$ under the true model $y_i \sim N(\theta, 1); \theta \sim N(0, 1)$ we compute a posterior sample of $T$ given each $y_r$. The posterior sample densities of these are plotted in light color. (Left) We simulate $y^{(true)}_{obs}$ under the true model and compute a posterior sample of $T$ given $y^{(true)}_{obs}$. Its sample density is plotted in dark color. (Right) We simulate $y^{(false)}_{obs}$ under the model $y_i \sim N(\theta, 1); \theta \sim N(3, 1)$ and compute a posterior sample of $T$ given $y^{(false)}_{obs}$. Its sample density is also plotted in dark color. We see that the posterior sample density of $T$ given $y^{(true)}_{obs}$ is very similar to those produced from the data replicated under the model. The sample density of $T$ given $y^{(false)}_{obs}$ is not very similar, being shifted to the right of the other sample densities.

### 1.2.3.2 Pivotal Discrepancy Measure (PDM) Method

Another model checking method was proposed by [Johnson (2007)] and [Yuan and Johnson (2011)]. We call this the Pivotal Discrepancy Measure (PDM) method. This method is based on using discrepancy measures that are pivotal quantities. Here, a pivotal quantity is a function $T(y, \theta)$ whose distribution, when evaluated at the true parameter $\theta_0$, does not depend on $\theta$. For example, from our model in (1.10) with $y_i \sim N(\theta, 1)$ then $T(y, \theta) = \bar{y} - \theta$ is a pivotal quantity because $\bar{y} \sim N(\theta, 1/n)$ and so $\bar{y} - \theta \sim N(0, 1/n)$, which is free of $\theta$.

[Johnson (2007)] and [Yuan and Johnson (2011)] prove some theoretical results that allow discrepancy measures that are pivotal to be used in Bayesian model checking. These results are discussed in detail in chapter 3 but to get the idea we will briefly describe them here. Suppose $y_{obs} \sim f(y|\theta_0)$ where $\theta_0$ represents the true data generating parameter value. It
is shown that if we simulate a posterior parameter value $\tilde{\theta} \sim \pi(\theta | y_{\text{obs}})$ from the posterior distribution of $\theta$ given $y_{\text{obs}}$, then the distributions of $T(y_{\text{obs}}, \theta_0)$ and $T(y_{\text{obs}}, \tilde{\theta})$ are identical. For example, suppose we again take $y_{\text{obs}} \sim N(\theta_0, 1)$ ($\theta_0$ unknown) and the pivotal quantity $T(y, \theta) = \bar{y} - \theta$. Then $T(y_{\text{obs}}, \theta_0)$ has a $N(0, 1/n)$ distribution, and if we draw $\tilde{\theta} \sim \pi(\theta | y_{\text{obs}})$ then we have that $T(y_{\text{obs}}, \tilde{\theta}) = \bar{y} - \tilde{\theta}$ has a $N(0, 1/n)$ distribution also.

This is a very useful result. Since $\theta_0$ is unknown we can not compute the value of $T(y_{\text{obs}}, \theta_0)$, but we can simulate a posterior value $\tilde{\theta}$ and compute a value of $T(y_{\text{obs}}, \tilde{\theta})$. The distribution of $T$ under the model can then be used as a reference distribution for this. [Johnson (2007)] and [Yuan and Johnson (2011)] recommend simulating a whole sample of posterior parameters $\tilde{\theta}_1, \ldots, \tilde{\theta}_N \sim \pi(\theta | y_{\text{obs}})$, and then computing a posterior sample $T(y_{\text{obs}}, \tilde{\theta}_1), \ldots, T(y_{\text{obs}}, \tilde{\theta}_N)$ of discrepancy measures. This posterior sample of the measure $T$ is then compared to the reference distribution of $T$.

[Johnson (2007)] gives a procedure for comparing the posterior sample of $T$ to the reference distribution, which will be discussed in chapter 3. Here we give a graphical example displayed in figure 1.5. In the top and bottom of figure 1.5 we plot histograms of the posterior samples of $T$ given $y_{\text{obs}}^{(\text{true})}$ and $y_{\text{obs}}^{(\text{false})}$, respectively. The density curve of the $N(0, 1/n)$ reference distribution of $T$ under the model is overlaid in each plot for reference. We see that the sample of $T$ given $y_{\text{obs}}^{(\text{true})}$ is similar to the reference distribution. The sample of $T$ given $y_{\text{obs}}^{(\text{false})}$ is shifted to the right, indicating the sample values are not typical of a sample from the reference distribution. This implies that the data $y_{\text{obs}}^{(\text{false})}$ was not generated under the model in (1.10).

1.3 Bayesian Model Checking Procedures Applied to Bayesian Item Response Theory Models

This dissertation will focus on applying the posterior predictive (PP), PPPS, and PDM Bayesian model checking procedures to Bayesian two parameter Item Response Theory mod-
Figure 1.5 (Top) Histogram of the posterior sample of $T$ given $y_{obs}^{(true)}$. (Bottom) Histogram of the posterior sample of $T$ given $y_{obs}^{(false)}$. The density curve of the $N(0, 1/n)$ reference distribution of $T$ under the model is overlaid for reference. We see that the sample of $T$ given $y_{obs}^{(true)}$ is similar to the reference distribution. The sample of $T$ given $y_{obs}^{(false)}$ is shifted to the right, indicating the sample values are not typical of a sample from the reference distribution.
els and testlet models. We will first give a brief introduction to IRT models. Then, we will discuss how we will apply these procedures to person fit checking. We will then introduce what a testlet model is and discuss how we will apply the PP and PPPS procedures for goodness-of-fit checking.

1.3.1 Brief Introduction to Item Response Theory Models

Item Response Theory (IRT) is a system of models used in the field of Psychometrics. Psychometrics is concerned with measuring psychological attributes of people such as knowledge level, ability, or intelligence. Such attributes, called latent traits, can not be measured directly because they exist as concepts and not physical dimensions. Exams or tests are given to people in which the observed responses can be used to learn about the latent trait(s) of interest. A basic example is the Stanford-Binett Intelligence Test which attempts to measure someone’s (unobservable) intelligence based on their (observable) test answers.

IRT models are a family of models used to analyze the responses of individuals to a test or exam. Exam questions are often referred to as items. Given these item responses, IRT models can be used to estimate the latent trait(s) on some scale. IRT models also allow for analyzing characteristics of the exam items, such as the difficulty level of the individual questions.

We will discuss the specifics of IRT models in chapter 1, but mention a few details here. There are several classifications for IRT models. With respect to the type of item responses, there are dichotomous models and polytomous models. In dichotomous models the responses are binary, having two possible outcomes. In this the responses are usually recorded as either correct or incorrect. Polytomous models allow for responses with several possible outcomes, e.g. multiple choice test items.

The parameters of the model are the latent trait parameter $\theta$, and the parameter(s) corresponding to each test item called item parameters. The latent trait parameter $\theta$ is believed to underlie the examinee’s performance in answering the test items. The item
parameters control characteristics of the items such as difficulty.

The dimensionality of an IRT model describes the number of latent traits that are believed to underlie examinee performance. In a unidimensional IRT model there is only one latent trait and $\theta$ is singular. In multidimensional IRT models there are believed to be multiple latent traits and so $\theta$ is a vector. For example, imagine a mathematics test where the questions rely entirely on a student’s mathematical ability for computation. Such questions might look like,

$$Q: \text{Solve the quadratic equation } x^2 + 2x - 14 = 1. \quad (1.13)$$

If a question were more like a word problem where an examinee must read a passage and then try to do some computation, one might argue that an ability for reading comprehension is needed along with the math ability.

Finally, IRT models can be described by the number of item parameters that are used for each item. A one-parameter (1P) model has only one parameter corresponding to each test item. Similarly, a two-parameter (2P) and three-parameter (3P) IRT model have two and three parameters, respectively, corresponding to each item.

We will be focusing on a unidimensional dichotomous two-parameter IRT model. Here we consider the response data $y$ to take values $y = 1$ if the item was answered correctly and $y = 0$ if answered incorrectly. There is one latent trait $\theta$, and two item parameters denoted $a$ and $b$. Here, $a$ is the item discrimination parameter and $b$ is the item difficulty parameter. Suppose that an exam is composed of $K$ items and is administered to $I$ examinees. Then the $k^{th}$ item, $k = 1, \ldots, K$, would would have a pair $a_k$ and $b_k$ corresponding to it, and the $i^{th}$ examinee, $i = 1, \ldots, I$, would have a latent trait parameter $\theta_i$ corresponding to them. The response data value $y_{ik}$ would then denote the response of examinee $i$ to item $k$.

IRT models attempt to model the probability of a correct response of an examinee. For our two-parameter model, the probability of examinee $i$ getting a correct response to item $k$
is given by,

\[ P(y_{ik} = 1|\theta_i, a_k, b_k) = F(a_k\theta_i - b_k), \]  

(1.14)

where \( F(.) \) is a link function. Similar to binary regression models, the function \( F \) is used to link the linear combination of \( a_k, \theta_i \), and \( b_k \) to a probability. Therefore, using cumulative distribution functions (cdf’s) is typical. One choice of \( F \) is the standard normal normal cdf, denoted by \( \Phi(.) \). So, our model can be specified by,

\[ P(y_{ik} = 1|\theta_i, a_k, b_k) = \Phi(a_k\theta_i - b_k). \]  

(1.15)

Because the standard normal cdf \( \Phi(.) \) is often referred to as the probit link, the model in (1.15) is sometimes referred to as a two-parameter probit IRT model. An IRT model using this link is also referred to as a normal ogive (pronounced o-jive) IRT model \cite{Lord and Novick 1968}. We will refer to the model in (1.15) as a two-parameter normal (2PN) IRT model. Another common choice of \( F \) is the standard logistic cdf. If this were used then the model would be a two-parameter logistic (2PL) IRT model. In this dissertation however, we will be focusing only on the normal model.

\subsection*{1.3.2 Model Checking in IRT Models}

Any statistical model makes certain assumptions about the data to which it applies. We have already mentioned one assumption that this 2PN IRT model makes. That is the assumption of dimensionality. In this model we are assuming their to be only one latent variable that underlies each examinees performance on the exam. Another assumption, which we will refer back to in a little while, is that of local independence. This is the assumption that, conditional on their ability parameter \( \theta_i \) and all item parameters \( a_k \) and \( b_k, k = 1, \ldots, K \), the item responses \( y_{i1}, \ldots, y_{iK} \) of examinee \( i \) are independent. It is common when conducting model diagnostics to check whether the modeling assumptions hold for the data. In IRT, dimensionality and local independence are typical assumptions to check.
We can also check the overall goodness-of-fit of a model to a set of data. In the whole process of model selection, a researcher may have several competing models to apply to a set of data. A preliminary step before attempting to choose a best fitting model is to check the goodness-of-fit of each model. Any model that is found to not fit the data well, may be eliminated from the pool. We will return to this later.

1.3.2.1 Research Focus 1: Person Fit Checking in IRT Models

Once a model has been selected it is common in statistical modeling to have a way to check for, or distinguish, data values that are outlying under the model. Assuming a model holds, outlying data values are usually values considered to be extreme or unlikely under the model. In IRT models there is a concept of outlying data values. Instead of single response data values \( y_{ik} \), we are concerned with assessing the fit of an examinee’s entire set of item responses \( y_i = (y_{i1}, \ldots, y_{iK})^t \), called a response pattern. This is referred to as person fit, or examinee fit, and is used to identify examinees whose responses are atypical or “outlying” under the model.

So, what would an outlying set of responses look like? This will be discussed in more detail in chapter 3. For now, response patterns that are produced under the model should look like Guttman patterns \(^1\text{Guttman}1950\). This means if all the exam items were ordered from easiest to hardest, the responses would have mostly correct answers at the beginning and incorrect answers at the end. For example, suppose a test had 10 questions and we rearranged an examinee’s responses so they started with the answers to the easiest questions and ended with the hardest. Then, under the model we would expect the response pattern to look something like (1) 1110110000 or (2) 1111011000, where 1 and 0 indicate a correct and incorrect response, respectively.

This makes sense, because depending on an examinee’s ability level we would generally expect that they would get the easiest questions correct. However, as the difficulty of the questions eventually exceeds their ability level, they will start having incorrect answers.
Examples of response patterns that would not follow in this manner would be something like (3) 0001001111 or (4) 0101010101. In (3) we see a pattern where most of the incorrect responses occur with the easiest items and correct responses occur with the hardest items. Also, in (4) the responses are correct/incorrect for easy and hard items alike.

Certain reasons have been hypothesized in the literature (Meijer 1996) as to why examinees may produce aberrant response patterns under a given IRT model. One is that an examinee may have had prior knowledge to the answers of some of the more difficult questions, i.e. the may have cheated. Another, is that an examinee may just have guessed their way through part, or all of the test. Incidentally, response patterns (3) and (4) above may resemble what responses produced from cheating or guessing, respectively, may look like.

Person fit checking in IRT then, is essentially trying to indicate those examinees whose response patterns significantly deviate from a Guttman-like pattern. Why is this of interest? Well, the problem with aberrant response patterns occurring is that the estimation of the examinees’ abilities $\theta_i$, $i = 1, \ldots, I$, is relative and dependent on the performance of all the examinees. It is important then to indicate those examinees whose response patterns are not in accordance with the model. That is, identify the “outlying” response patterns. After doing this, researchers may choose to remove those examinees’ responses and re-fit the model.

In a classical setting, there are many statistics that are used to analyze person fit. Meijer and Sijtsma (2001) give an overview of the most commonly used person fit statistics. For example, a commonly used person fit statistic is a residual based statistic introduced by Wright (1980). For a given set of responses $y_{i1}, \ldots, y_{iK}$, under the model each $y_{ik}$ is a Bernoulli random variable ($y_{ik} = 1$ or 0). Letting $p_{ik}$ denote the probability of success, then the statistic of Wright (1980) is given by,

$$W_i = \frac{\sum_{k=1}^{K} (y_{ik} - p_{ik})^2}{\sum_{k=1}^{K} p_{ik}(1 - p_{ik})}. \quad (1.16)$$
If we are assuming a 2PN IRT model as in (1.15), then $p_{ik} = \Phi(a_k\theta_i - b_k)$, $k = 1, \ldots, K$. In a classical setting $W_i$ is computed by estimating $p_{ik}$ by $\hat{p}_{ik} = \Phi(\hat{a}_k\hat{\theta}_i - \hat{b}_k)$, where $\hat{a}_k$, $\hat{\theta}_i$, and $\hat{b}_k$ are the maximum likelihood estimates of $a_k$, $\theta_i$, and $b_k$, respectively. From this, person fit would be assessed by taking the computed value of $W_i$, $\hat{W}_i$, and comparing it to its sampling distribution. If $\hat{W}_i$ was extreme within its sampling distribution, this would suggest that the response pattern $y_{i1}, \ldots, y_{iK}$ was not in agreement with the model.

These statistics can be adopted for use in a Bayesian setting. In a Bayesian setting, given the observed data we can simulate posterior values of the parameters $a_k$, $b_k$, and $\theta_i$, denoted $\tilde{a}_k$, $\tilde{b}_k$, and $\tilde{\theta}_i$, from their respective marginal posterior distributions. With these we can compute a posterior value of $p_{ik}$, namely $\tilde{p}_{ik} = \Phi(\tilde{a}_k\tilde{\theta}_i - \tilde{b}_k)$. So, we can compute a Bayesian “counterpart” of the statistic $W_i$ in (1.16) using $\tilde{p}_{ik}$, which will be a posterior value of $W_i$. In practice we can simulate many posterior parameter values $\tilde{a}_k^{(n)}$, $\tilde{b}_k^{(n)}$, and $\tilde{\theta}_i^{(n)}$, $n = 1, \ldots, N$, which we use to compute $\tilde{p}_{ik}^{(n)} = \Phi(\tilde{a}_k^{(n)}\tilde{\theta}_i^{(n)} - \tilde{b}_k^{(n)})$, $n = 1, \ldots, N$. These in turn can be used to compute a posterior sample of $W_i$ given by,

$$\tilde{W}_i^{(n)} = \sum_{k=1}^{K} \frac{(y_{ik} - \tilde{p}_{ik}^{(n)})^2}{\sum_{k=1}^{K} \tilde{p}_{ik}^{(n)}(1 - \tilde{p}_{ik}^{(n)})},$$

$n = 1, \ldots, N$. This posterior sample of $W_i$ given the observed data can now be used under each of the three methods: PP, PPPS, or PDM. For example, under the PP method this posterior sample can be compared to simulated values of $W_i$ under the posterior predictive distribution to assess person fit. If the posterior values $\tilde{W}_i^{(n)}$ are much larger (or smaller) in comparison to the posterior predictive values, this would indicate the response pattern $y_{i1}, \ldots, y_{iK}$ was not in agreement with the model.

The research conducted on person fit analysis in Bayesian IRT models has been somewhat limited. Using the PP method, Glas and Meijer (2003) conducted a simulation study investigating the performance of the $W$ statistic among other “classical” type person fit statistics in detecting person misfit in a three-parameter Normal Bayesian IRT model. Also
using the PP method, Toribio (2006) conducted similar simulation studies using some of the same person fit measures as Glas and Meijer (2003), to assess person fit in a two-parameter Normal Bayesian IRT model.

Among the research conducted into person fit in Bayesian IRT models so far, the model checking method employed has been the PP method. Our research interest is to see if applying the two new model checking measures, PPPS and PDM, will be more effective in assessing person fit. Of course, assessing the “effectiveness of the method” is somewhat vague since it will come down to the individual performance of the person fit measures under each method. We will look at using many of the same person fit measures used by Glas and Meijer (2003) as well as a few others (these will be introduced in chapter 4). Most of the person fit measures we will use will be classical measures such as $W$, that will be implemented in a Bayesian setting (i.e. under the PP, PPPS, and PDM methods). We will also define a new person fit measure based on Bayesian latent residuals (Albert and Chib 1995). As we will see in chapter 4, this measure will be completely defined as a function of the model parameters only and can be computed using simulated posterior parameter values.

The person fit measures will be used to assess person fit of response data fit under a Bayesian 2PN IRT model using each of the three model checking methods, PP, PPPS, and PDM. Specifically, using each method we will conduct a simulation study where 100 separate response data sets $Y_{obs}$ will be generated under the model, but with certain examinees having response patterns simulated to be aberrant under the model. For each data set, we will evaluate the person fit measures (under the respective method; PP, PPPS, or PDM) and record how many of the violating examinees each measure indicated. Under each model checking method, we will be able to look at the number of times a particular measure correctly identified an examinee’s aberrant response pattern.

A goal of ours is to see under which method each of the person fit measures is most effective in detecting aberrant response patterns. That is, over the 100 data sets $Y_{obs}$ we want to see under which method each measure detected the highest number of aberrant
response patterns. From here, we can answer two questions: (1) given a particular person fit measure, under which method is it most effective, and (2) for a particular method, which measure performs best under it? So, if a researcher were to perform person fit checking and had a particular measure in mind, they would know the best method to use it under. Likewise, if they were going to employ a particular model checking method, they would know the most effective measure(s) to use under it. We will also look at the overall highest detection rates, and see under which method (and by which measure) these were obtained. This will tell the overall best measure/method pair to use. A particular interest is to see if these occur under either of the two new methods: PPPS or PDM.

We will also investigate the type I error rates of the measures, used under the different methods. In this we will generate 100 data sets $Y_{obs}$ with all examinee response patterns simulated under the model. For each data set, we will evaluate the person fit measures under the three methods and record how many of the violating examinees each measure indicated. We will be able to see the number of times each measure incorrectly indicate a lack of fit of an examinee’s response pattern under the model.

The questions we want to answer are:

1. What are the approximate type I error rates of each person fit measure, used under its respective method?

2. For each person fit measure, under which method (PP, PPPS, PDM) is it most effective in detecting a lack of fit with the examinee response patterns simulated to be aberrant under the model?

3. Under each method, which are the best performing measures in detecting a lack of fit with the examinee response patterns simulated to be aberrant under the model?

4. Looking at the overall highest detection rates in the simulation study, which measure, and under which method, produced these? Were these produced by a measure used under either of the new PPPS or PDM methods?
1.3.2.2 Research Focus 2: Goodness-of-Fit Checking for 2PN Testlet Models

Another type of model in IRT is a testlet model (Bradlow et al. 1999, Wainer et al. 2007). A testlet (Wainer and Kiely 1987) is a group of items on an exam that are based on a common stimulus. For example, a word problem on a mathematics exam may have multiple questions relating to it. All items relating to the passage make up a testlet. We note that there may be multiple testlets on an exam. A result of having multiple items related like this is that there may be some dependence among examinees’ responses to them. So, if an exam has testlets then the assumption of local independence does not hold. A testlet model is basically an IRT model with an added parameter that tries to account for the dependence of item responses within a testlet.

We will discuss the specifics of testlet models in more detail in chapter 5, but give a brief example so the reader has an idea of the difference between these and the IRT models previously discussed. For example, the standard 2PN IRT model we saw before in (1.15) was

\[ P(y_{ik} = 1 | \theta_i, a_k, b_k) = \Phi(a_k \theta_i - b_k). \]

Now, suppose that for an exam with \( K \) items, the first eight items make up a testlet. Then there is some dependence among these item responses and we model this with the parameter \( \gamma \). If we let \( A = \{1, 2, \ldots, 8\} \), a testlet model based on the 2PN IRT model above could be given as,

\[ P(y_{ik} = 1 | \theta_i, a_k, b_k, \gamma) = \Phi(a_k \theta_i - b_k - I_{(k \in A)} \cdot \gamma), \quad (1.18) \]

where \( I_{(k \in A)} \) is an indicator function in which \( I_{(k \in A)} = 1 \) if \( k \in A \), and 0 otherwise. So, for all items in the testlet (i.e. for \( k \in A \)) the model has the extra parameter \( \gamma \) and the probability of a correct response to those items for examinee \( i \) is \( \Phi(a_k \theta_i - b_k - \gamma) \). For all other items the probability of a correct response is just \( \Phi(a_k \theta_i - b_k) \).

We discuss some issues that occur when the 2PN IRT model in (1.15) is fit to data.
actually generated under a 2PN testlet model. In practice, it may be known that some
test items may rely on common stimuli, yielding a dependence among the the responses.
However, this dependence is often ignored to yield a more straightforward analysis (Bradlow
et al. 1999). This means that even though testlets may be seen to exist in an exam, the
assumption of local independence is still made and standard IRT models such as in (1.15)
are used instead of a testlet model. This is because standard IRT models such as in (1.15)
are more commonly used in test analysis. Also, these models are somewhat simpler due to
the lower number of parameters.

Problems do exists when the dependence in testlets is ignored and a standard IRT model
such as in (1.15) is applied. A result, as mentioned in Bradlow et al. (1999), is that there
will be bias in the posterior item parameter estimates of $a_k$ and $b_k$, $k = 1, \ldots, K$, obtained
from fitting the model. However, Bradlow et al. (1999) show that for data generated under
a testlet model, if the dependence among the testlet items is not very high then a standard
IRT model may fit the data well. So, having a way to check the goodness-of-fit of a standard
IRT model to such a data set would be useful.

Research has been conducted by Sinharay and Johnson (2003) in checking the goodness-
of-fit of a three-parameter logistic (3PL) IRT model when fit to data generated under a 3PL
testlet model. Using the PP method, Sinharay and Johnson (2003) used test statistics based
on biserial correlations as goodness-of-fit measures. Briefly, a *biserial correlation coefficient*
(Pearson 1909) measures the correlation between a dichotomous variable $Y$ (e.g. $Y = 0$ or
1) and a continuous variable. For IRT response data, the sample biserial correlation can be
computed over all examinee responses $y_{1k}, \ldots, y_{Ik}$ to each item $k$, for $k = 1, \ldots, K$. This
gives a total of $K$ computed biserial correlations, one for each item. In practice, this can
be used to show the strength of the relationship between the item responses $y_{1k}, \ldots, y_{Ik}$
and the latent trait $\theta$, whose scale is assumed to be continuous. We will discuss these in
detail in chapter 4, but note for now that these biserial correlations are related to the item
discrimination parameters, $a_k$, $k = 1, \ldots, K$. 
For a set of observed response data $Y_{obs}$, Sinharay and Johnson (2003) used the sample mean, standard deviation, and maximum of the $K$ computed biserial correlations as goodness-of-fit measures under the PP method. The results showed that these measures were not very effective in detecting a lack of fit of the 3PL IRT model to 3PL testlet model data. In this dissertation we will focus on the two-parameter normal case. That is, checking the goodness-of-fit of the 2PN IRT model to response data generated under a 2PN testlet model.

Sinharay and Johnson (2003) conducted their research using the PP method. We will look at applying the PP and PPPS model checking methods to check the goodness-of-fit of the 2PN IRT model to 2PN testlet model response data. Under the PP method we will use same three measures used by Sinharay and Johnson (2003): the mean, standard deviation, and maximum of the sample biserial correlations. Under the PPPS method we will define three new model checking measures to use. These will be based on simulated posterior values of the item discrimination parameters $a_k$, $k = 1, \ldots, K$. Specifically, we will use the mean, standard deviation, and maximum of posterior samples of $a_k$, $k = 1, \ldots, K$. The motivation for this, which will be explained in detail in chapter 5, ties into the previously mentioned occurrence of bias in the posterior estimates of these parameters when the (wrong) 2PN IRT model is fitted.

We are interested in seeing how effective these chosen measures under their respective methods are in indicating a lack of fit of a 2PN IRT model to testlet exam data. To do this we will conduct two simulation studies. One will investigate the measures under the PP method and the other will investigate the measures used under the PPPS method. For each iteration in the simulation study will be checking the goodness-of-fit of the 2PN IRT model to four “observed” testlet data sets $Y_{obs}$. The four data sets will be generated such that there are varying levels of dependence within the testlets. The simulation details will be fully explained in chapter 4, but for now the levels of dependence will very from very weak to fairly strong. This will be to see how well the measures, under the respective methods,
detect a lack of fit when the dependence is weaker versus stronger.

To measure “how well” each measure does in detecting a lack of fit of the 2PN IRT model, we iterate the process 100 times (for each of the 4 observed data sets mentioned above). We then look at the number of times out of 100 simulations that each measure indicated a lack of fit.

We also investigate the type I error rates of the measures used under the PP and PPPS methods. In this, we will again simulate four observed data sets under the 2PN testlet model with the same levels of dependence within testlets as above. We will then fit the true data generating model to each data set. We iterate this 100 times and record the number of times each measure indicated a lack of fit. This gives us an idea of how often each measure, used under its respective method, may wrongly indicate a lack of fit with the true model.

The questions we want to answer are:

1. What are the approximate type I error rates of each measure, used under its respective method.

2. Under each method, which measure performs best in detecting the lack of fit of the 2PN IRT model to response data generated under the 2PN testlet model? In particular, we are interested in seeing if the measures under the PPPS method outperform those under the PP method. Since the PP measures we will use did not show to be that effective in the simulation study done by Sinharay and Johnson (2003), we want to know if any of the new measures used under the PPPS method are an improvement.

3. How strong does the dependence of the item responses within the testlets need to be in order for the measures to be effective?
CHAPTER 2
INTRODUCTION TO ITEM RESPONSE THEORY MODELS

2.1 Introduction to Item Response Theory

Item Response Theory (IRT) is a system of modeling used in the field of psychometrics. Psychometrics is concerned with the theory and practices of measuring psychological attributes of people like knowledge level or ability. These attributes can not be measured directly such as a person’s height or weight because they exist as concepts and not physical dimensions. The term that is used to describe such unobservable attributes is latent traits.

IRT models focus on trying to measure these latent traits through other concrete observations. A major area that IRT models are applied is testing in education. Here, an examinee’s responses to questions can be used to estimate the ability of the student in a given subject. Just as the Stanford-Binet Intelligence Test attempts to quantify someone’s intelligence on an intelligence quotient (IQ) scale, similarly IRT attempts to quantify a person’s given subject ability via this latent trait. For example, the Graduate Record Exam (GRE) has a quantitative section that asks students to answer mathematically related questions. The latent trait that examiners want to measure is a person’s mathematical ability.

In IRT the latent trait, or variable, that is measured is referred to as the ability parameter, typically denoted by \( \theta \). The questions on a testing device are referred to as items and it is the examinees’ responses to these items that are believed to be dependent on their ability \( \theta \). A main endeavor in IRT is to estimate these parameters. This can allow examiners to measure the abilities of the examinees and also to analyze the items on the exams they create.

The purpose of the rest of this chapter is to give a general introduction to IRT models and supply the reader with the necessary prerequisite knowledge to understand the IRT models.
used in subsequent chapters. We will begin by introducing IRT models from a frequentist standpoint, discussing their main characteristics. After this we will give an introduction to Bayesian modeling and then discuss IRT models in a Bayesian setting. We will then look at IRT parameter estimation techniques under both a frequentist and Bayesian approach, and then compare these. We finish the chapter with some motivations as to why we prefer to use Bayesian IRT models as opposed to frequentist models.

### 2.2 Item Response Models for Binary Responses

In item response models we consider the response of an examinee to be binary, that is it is either correct or incorrect. We let $y$ denote an examinee’s response to an item which takes a value of $y = 1$ if the item is answered correctly and $y = 0$ if it is answered incorrectly. IRT models then model the probability of a correct response, or that the random variable $y$ equals one.

In this, IRT models are similar to binary regression models which attempt to model the probability of success $y = 1$, or failure $y = 0$, of a random event. In a binary regression model, the probability of success is given by,

$$P(y = 1) = F(X\beta)$$

$$= F(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n),$$

where $X$ is a covariate matrix, $\beta$ is an unknown parameter vector, and $F$ is a known link function. A natural choice of such functions $F$ are cumulative distribution functions (cdf’s) because they link the linear predictor $\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$ to the probability of success by mapping it to the interval $(0, 1)$. Some typical choices of $F$ are the standard logistic cdf given by $\Psi(x) = e^x/(1 + e^x)$, and the standard normal cdf which we denote by $\Phi(.)$. Binary regression models that use the logistic link function are called logistic models and models that use the standard normal link function are called probit models. As we will soon see,
IRT models also employ these link functions and in fact inherit similar names due to that. We will next introduce and discuss three commonly used IRT models.

In IRT models the probability of an examinee’s correct response to an item is a function of their underlying ability $\theta$. IRT models are described by the number of item parameters that are used for each item. For example, a one-parameter IRT model has only one parameter corresponding to each item. Similarly, a two-parameter and three-parameter IRT model have two parameters and three parameters, respectively, corresponding to each item. Next we will look at each of these three types of IRT models. We will mention which type of item parameters are used in each model, but describe them in more detail later on.

Suppose we have a test with $K$ items. For each item $k = 1, \ldots, K$ on the test, a one parameter IRT model would have only an item difficulty parameter, $b_k$. A two parameter IRT model would have both an item difficulty parameter $b_k$ and an item discrimination parameter, $a_k$ for each item. The three parameter model has the discrimination and difficulty parameters, but also has a guessing parameter $c_k$ for each item. The common notation for IRT models with one, two, or three parameters is 1PN, 2PN, 3PN, for the normal ogive model and 1PL, 2PL, 3PL for the logistic model.

### 2.2.1 One Parameter Model

In a one-parameter IRT model the only item parameter is the difficulty parameter $b$. In this one parameter case, the probability of an examinee answering an item correctly is a function of the difference between the examinee ability parameter and the item difficulty parameter. The probability of answering an item correctly is modeled as

$$P(y = 1|\theta, b) = F(\theta - b). \quad (2.2)$$

As mentioned, the link function $F$ is typically taken to be the standard logistic or standard normal cdf. The two models are shown below.
1. One-Parameter Logistic (Rasch) Model

In this model, \( F(.) = \Psi(.) \) and it is referred to as the one-parameter logistic (1PL) model. The 1PL model is given by,

\[
P(y = 1|\theta, b) = \Psi(\theta - b) = \frac{e^{\theta - b}}{1 + e^{\theta - b}}.
\]  

(2.3)

This particular model is also referred to as the Rasch model \(^{(Rasch 1966)}\).

2. One-Parameter Normal Model

In this model, \( F(.) = \Phi(.) \) and it is referred to as the one-parameter probit model. The 1PN model is given by,

\[
P(y = 1|\theta, b) = \Phi(\theta - b).
\]  

(2.4)

This model is also referred to as the one-parameter normal ogive (1PN) model.

2.2.2 Two-Parameter Model

In the two-parameter item response model there are two item specific parameters, \( a \) and \( b \) which describe the item. Here, \( a \) is the item \textit{discrimination} parameter and \( b \) is the item \textit{difficulty} parameter. With these, along with an examinee’s ability parameter \( \theta \), the probability of answering an item correctly is modeled as

\[
P(y = 1|\theta, a, b) = F(a\theta - b).
\]  

(2.5)

We can see that the one-parameter model is a special case of the two-parameter model with discrimination parameter \( a \) taken to be constant at 1. In \( \text{[2.5]} \), using the logistic or probit link for \( F \), gives the two-parameter logistic (2PL) or two-parameter normal ogive (2PN) models, respectively. These two models are shown below.
1. Two-Parameter Logistic (2PL) Model

\[ P(y = 1 | \theta, a, b) = \Psi(a \theta - b) = \frac{e^{a \theta - b}}{1 + e^{a \theta - b}}. \] (2.6)

2. Two-Parameter Normal (2PN) Model

\[ P(y = 1 | \theta, a, b) = \Phi(a \theta - b). \] (2.7)

It is worth mentioning that for the 1P and 2P models there is a scale factor \( d \) that is commonly used to obtain an approximate equivalence between the standard normal and logistic distributions. The factor constant used is \( d = 1.7 \), and with this we have

\[ \Phi(x) \approx \Psi(d \cdot x). \] (2.8)

In particular, for all real values of \( x \) we have that,

\[ |\Phi(x) - \Psi(d \cdot x)| < 0.01. \] (2.9)

A background of the factor \( d \) can be found in Camilli [1994]. For the IRT models discussed so far the scale factor \( d \) can be applied to give,

\[ P(y = 1 | \theta, a, b) = \Phi(a \theta - b) \approx \Psi(d(a \theta - b)). \] (2.10)

It should be noted here that there is an underlying assumption of the two aforementioned models. This assumption is that while taking the exam, all examinees are making a genuine effort in answering every item to the best of their ability. What the above models do not take into account is the possibility of an examinee getting an item correct by guessing at it.
2.2.3 Three Parameter Model

What the three-parameter IRT model takes into consideration is the possibility that examinees can get an answer to an item correct by guessing. Therefore a guessing parameter \( c \) is included. The probability of answering an item correctly is given by,

\[
P(y = 1|\theta, a, b) = c + (1 - c)F(a\theta - b).
\]  

(2.11)

The parameter \( c \) represents the probability of an examinee getting an item correct by only guessing. This parameter has a theoretical range of \( 0 \leq c \leq 1 \), but values above 0.35 are generally not considered acceptable (Baker 2001). An example of where this model might be applied is in a multiple choice setting. Suppose that an item is a multiple choice question with 4 possible answers to select from. If an examinee is to guess the answer then theoretically they have a 1 in 4 or 25% chance of getting it correct. In this case, the value of \( c \) would be set to 0.25. However, from (2.11) we see that 0.25 would not be the probability of getting the item correct. The three-parameter model considers guessing as contributing to the overall probability of success.

As with the previous models, using the logistic or probit link for \( F \) in (2.11) gives the three-parameter logistic (3PL) or three-parameter normal ogive (3PN) models, respectively. These are given below.

1. Three-Parameter Logistic (3PL) Model

Birnbaum (1968) modified the 2PL model to include the guessing parameter \( c \). The 3PL model is given by,

\[
P(y = 1|\theta, a, b, c) = c + (1 - c)\Psi(a\theta - b)
\]  

(2.12)

\[
= c + (1 - c)\frac{e^{a\theta - b}}{1 + e^{a\theta - b}}.
\]
2. Three-Parameter Normal (3PN) Model

\[
P(y = 1|\theta, a, b, c) = c + (1 - c)\Phi(a\theta - b). \tag{2.13}
\]

2.2.4 Model Assumptions

The above set of IRT models have a common pair of assumptions that we will mention before continuing. The first assumption deals with the \textit{dimensionality} of the model. Dimensionality refers to the number of latent traits or abilities that underlie an examinee’s performance in answering the items on an exam. The aforementioned models are \textit{unidimensional}, in that they all assume only one latent trait is needed to account for an examinee’s performance. That is, an examinee’s ability \(\theta\) is a singleton. There are \textit{multidimensional} IRT models that assume more than one ability is needed to account for an examinee’s performance. Here, \(\theta\) is a vector whose elements account for the different abilities. These are an interesting group of models also, but for this dissertation work we will be focusing on models with the unidimensionality assumption.

The other assumption that the above models have is that of \textit{local independence}. This assumption is that, conditional on an examinee’s ability parameter and all item parameters, their test item responses are statistically independent. For example, suppose an exam with \(K\) items and we are considering a two-parameter model. Then we have \(K\) item discrimination and difficulty parameters denoted \(a = (a_1, \ldots, a_K)^t\) and \(b = (b_1, \ldots, b_K)^t\). For a given examinee with ability parameter \(\theta\), the local independence assumption can be expressed as,

\[
P(y_1 = 1, \ldots, y_K = 1|\theta, a, b) = \prod_{k=1}^{K} P(y_k = 1|\theta, a, b). \tag{2.14}
\]

A third modeling assumption is technically made for IRT models, and that is the parametric
form of the right side of (2.2), (2.5), and (2.11). This is sometimes referred to as the item
response function (IRF). We have already done this though by considering the logistic and
probit links. Next we will discuss another important aspect of the above models, which is
the item response curve. In our discussion of this we will also discuss the item parameters
in more detail.

2.3 Item Response Curve

The item response curve (IRC) is the plot of the item response function (IRF) given in
the right sides of (2.2), (2.5), and (2.11). Considering the logistic or probit link functions,
we can have IRC’s for each of the IRF’s in the right sides of (2.3), (2.4), (2.6), (2.7), (2.12),
or (2.13). For each item $k$, and corresponding set of item parameters, an IRC plots the
probability of a correct response as a function of the ability $\theta$. In an IRC for a given item,
the item parameters are considered to be fixed. For example, under the 2PN model for a
given item the IRC would be the plot of the IRF $P(\theta) = \Phi(a\theta - b)$, which is a function of $\theta$
with item parameters $a$ and $b$ fixed. It should be noted that we are assuming that the latent
ability parameter $\theta$, and the individual item parameters, are continuous and one-dimensional.

Theoretically, the latent ability parameter can assume any value on the real number line.
However, the scale used for the ability $\theta$ is determined by the assumptions of the model. In
the 2PN model for example, the probability of success for a given item is $\Phi(a\theta - b)$ where
it is assumed $\theta \sim N(0, 1)$ (Lord and Novick 1968 p.366). So, it is meaningful to think of $\theta$
taking values between about $-3$ and $3$ (which are inherited from the standard normal scale),
with $\theta = -3$ representing a very poor ability level, $\theta = 0$ representing an average ability level
and $\theta = 3$ a very good ability level. We generally assume that an examinee with a higher
ability level should do better in answering an item than an examinee with a lower ability
level. Due to this, we assume the probability of an examinee correctly answering an item to
be an increasing function of the ability $\theta$. This is illustrated in the example IRC plotted in
figure 2.1 where we plot a 2PN model with discrimination $a = 1$ and difficulty $b = 0$. 
Figure 2.1 An example item response curve using the 2PN model with discrimination $a = 1$ and difficulty $b = 0$.

It was stated above that $a$ is the item discrimination parameter. So, $a$ is the parameter that indicates how well the item discriminates between different ability levels of examinees. In reference to an IRC, $a$ is the parameter that determines the slope of the curve. The discrimination parameter $a$ is proportional to the slope of the IRC at its inflection point (De Ayala 2009). Look at figure 2.2 where we have three item response curves plotted using the two-parameter normal ogive model in (2.7). In each, we set the value of difficulty parameter, $b$, to 0. Each curve is plotted using a different value of the discrimination parameter $a$, namely $a = 0.4$, $1$, and $2$. We see that as the value of $a$ increases the curve becomes steeper. This reflects that with a larger value of $a$, the item discriminates between examinees of different ability levels better. For example, the curve with $a = 2$ has a rather steep slope. This means that an examinee’s probability of correctly answering the item increases very fast as their ability increases. So, this item discriminates very well between a good student and one who is not so good. The curve with $a = 0.4$ has a very shallow slope. So, the probability of correctly answering the item does not increase much as examinee ability level increases. This item would not discriminate as well between examinees with lower ability and higher ability. The parameter $b$ is the item difficulty parameter. This controls the location of
Figure 2.2 Three item response curves under a 2PN model with the same difficulty of $b = 0$, and different discrimination values.

Figure 2.3 Three item response curves with the same discrimination value of $a = 1$, and different difficulty values. A horizontal line is plotted at probability of success equal to 0.5.
the two-parameter IRC. The value $b$ on the ability scale is the point that corresponds to a 0.5 probability of correctly answering the item. An item is deemed to be more difficult if the probability of success is lower in comparison with another item given the same ability level. In figure 2.3 three more item response curves are plotted using the 2PN model as above, each with a different value of $b$. The values of $b$ are $-1$, $0$, and $1$. The values of $a$ are kept the same for all three curves with $a = 1$. The items of these curves are of increasing difficulty. Looking at the curve with $b = -1$, the probability of success if $\theta = -1$ is 0.5. For this same value of $\theta$ we see that the probability of success for the item with $b = 0$ is about 0.16, and about 0.02 for the item with $b = 1$. Another way to see that these items increase in difficulty is to notice that in order to maintain a probability of success 0.5 on each item the ability level has to increase from the left to the rightmost curve.

Theoretically, the item parameters $a$ and $b$ can take any value on the real line. Since we want to model the probability of an examinee’s correct response as an increasing function of their ability, the value of the discrimination parameter $a$ is taken to be positive. A reasonably “good” value of $a$ should yield an item that is helpful in discriminating subjects of different abilities. Such values of $a$ range from 0.8 to 2.5 under the logistic model, and dividing these by the scale factor $d = 1.7$ mentioned at the end of section 2.2.2 gives an equivalent range of approximately 0.5 to 1.5 under the normal ogive model (De Ayala 2009, p.101). A more detailed verbal scale for $a$ is given in Baker (2001, p.35).

Such a scale is not given for the difficulty parameter, but we can still get a relative idea if we consider the ability scale. It was mentioned that the value $b$ on the ability scale is the point that corresponds to a 0.50 probability of correctly answering the item. So, for difficulty parameter $b$, one needs an ability of $\theta = b$ to have a 50% chance of getting the item correct. We can consider a difficulty parameter value of $b = 0$ as being moderately difficult, because a person of average ability, $\theta = 0$, has a 0.50 probability of getting it correct. From this point the farther $b$ is in the negative direction the “easier” the item is, because one needs a lower and lower ability to have a 50% chance of correctly answering it. Likewise, the farther
$b$ is in the positive direction the “harder” an item gets.

An IRC for a three-parameter model looks a little different from that of the one and two-parameter models. We note that the two-parameter model is just a special case of the three-parameter model with the guessing parameter $c = 0$. For a three-parameter model with non-zero guessing parameter $c$ we have that the lower limit of the curve is $c$ and not zero. We see this in figure 2.4 where we plot three 3PN models with discrimination parameters equal to 1 and difficulty parameters equal to 0, but with guessing parameters equal to 0.10, 0.25, and 0.50, respectively. Recall, the value of $c$ is taken as the probability an examinee would correctly answer an item by just guessing. So, a plausible setting for these values would be a 10 part multiple choice item, a 4 part multiple choice item, and possibly a True/False item.

It was mentioned above that the three-parameter model considers guessing as contributing to the overall probability of success on an item. Looking at the curves in figure 2.4 we see that the lower an examinee’s ability is, and hence the lower their probability of success, the probability of correctly answering the item turns into the probability of correctly answering by just guessing. The definition of the difficulty parameter also changes. In the 3P case, the difficulty parameter is the point on the ability scale that gives probability of success that is midway between the lower limit $c$ and 1.

### 2.4 Introduction to Bayesian Methodology

#### 2.4.1 Introduction

Let $\omega$ be the vector of unknown parameters of interest. Bayesian methodology starts with assigning a prior distribution to $\omega$, which we will denote by $p(\omega)$. Specifying the prior distribution is the point where the researcher can quantitatively summarize any knowledge about $\omega$ prior to observing any data. This is a subjective probability distribution formulated based on the researcher’s prior beliefs and/or knowledge. Next, data $y$ is observed and is considered to have sampling density $f(y|\omega)$. As a function of $\omega$, $f(y|\omega)$ is the likelihood
What has happened is that our original prior knowledge is added to, or updated, when we observe data \( y \). The prior distribution and likelihood are used to obtain an updated posterior distribution of \( \omega \) which is conditional on the observed data \( y \). We denote this posterior distribution by \( \pi(\omega|y) \). This posterior distribution of \( \omega \) is mathematically obtained via Bayes Theorem. A simple form of Bayes Theorem, for events \( A \) and \( B \), with \( P(B) \neq 0 \) is given as

\[
P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \propto P(B|A)P(A). \tag{2.15}
\]

In terms of \( \omega \) and \( y \) we apply Bayes Theorem to get

\[
\pi(\omega|y) = \frac{p(\omega, y)}{f(y)} = \frac{f(y|\omega)p(\omega)}{f(y)}. \tag{2.16}
\]

Here, \( f(y) = \int f(y|\omega)p(\omega)d\omega \) is the marginal distribution of \( y \). This is the distribution of the data that we get by integrating out \( \omega \) from the joint distribution \( p(\omega, y) \). Since the
marginal distribution \( f(y) \) is free of the unknown parameters \( \omega \), then given observed data \( y \), \( f(y) \) can be thought of as a constant. Thus, from (2.16) we have

\[
\pi(\omega|y) \propto f(y|\omega)p(\omega). \tag{2.17}
\]

The posterior distribution of \( \omega \) given \( y \) is proportional to the likelihood times the prior distribution of \( \omega \). The marginal \( f(y) \) is viewed as a normalizing constant and omitting it will not affect the inferences made from the posterior distribution.

### 2.4.2 Bayesian IRT Model

Here we introduce the Bayesian IRT Model. For ease of explanation we will use the two-parameter model. In the following subsections 2.4.2.1, 2.4.2.2 and 2.4.2.3 we will specify the likelihood function, prior distributions and joint posterior distribution for the model parameters, respectively. In these subsections we are describing a general two-parameter Bayesian IRT model. In this, we will use \( F \) to denote the link function as before. In each of the following equations, \( F \) could be set equal to the logistic or probit links to give either a 2PL or 2PN Bayesian IRT model, respectively. After discussing the general form of two-parameter Bayesian IRT model in this section, we will discuss the 2PN Bayesian IRT model in more detail in the next section. Before all this, we first need to describe some specifics pertaining to the response data.

So far, we have considered a model for the response of an individual to a single test item. In general, we consider having \( I \) examinees take a test consisting of \( K \) items. We let \( y_{ik} \) denote the response of the \( i^{th} \) examinee to the \( k^{th} \) item on the test, coded with 1 for a correct response and 0 for an incorrect response. From this, the response data is organized into an \( I \times K \) matrix where the \( i^{th} \) row entries, \( y_{i1}, \ldots, y_{iK} \) contain the responses of the \( i^{th} \) examinee to the \( K \) test items. Likewise, the \( k^{th} \) column entries \( y_{1k}, \ldots, y_{Ik} \) contain the responses of the \( I \) examinees to the \( k^{th} \) item.
### 2.4.2.1 Likelihood Function

Suppose a test containing $K$ items is given to a group of $I$ examinees. Let $y_{i1}, y_{i2}, \ldots, y_{iK}$ be the binary responses of the $i$th examinee to the $K$ items and let $\mathbf{a} = (a_1, \ldots, a_K)$ and $\mathbf{b} = (b_1, \ldots, b_K)$ be the vectors of item discrimination and difficulty parameters, respectively. One assumption that we make is that the responses to each item by an individual are *locally independent*. Local independence is a term in IRT that means we assume an examinee’s responses to test items are conditionally independent given their ability parameter and all item parameters. Then the responses $y_{i1}, y_{i2}, \ldots, y_{iK}$, given $\theta_i$, $\mathbf{a}$, and $\mathbf{b}$, are a set of $K$ independent Bernoulli random variables. It follows from section 2.2.2 that the probability mass function of $y_{ik}$ can be written as

$$P(Y_{ik} = y_{ik} | \theta_i, a_k, b_k) = F(a_k \theta_i - b_k)^{y_{ik}} [1 - F(a_k \theta_i - b_k)^{(1-y_{ik})}],$$

(2.18)

for $y_{ik} = 0$, or 1. It follows that the probability of observing the $i$th examinee’s entire sequence of responses is given by the joint pmf

$$P(Y_{i1} = y_{i1}, \ldots, Y_{iK} = y_{iK} | \theta_i, \mathbf{a}, \mathbf{b}) = \prod_{k=1}^{K} P(Y_{ik} = y_{ik} | \theta_i, \mathbf{a}, \mathbf{b})$$

(2.19)

$$= \prod_{k=1}^{K} F(a_k \theta_i - b_k)^{y_{ik}} [1 - F(a_k \theta_i - b_k)]^{(1-y_{ik})}$$

Now, we also assume the responses among the $I$ examinees to the test items are independent. Then the likelihood function for all responses of all examinees is

$$L(\theta, \mathbf{a}, \mathbf{b}) = \prod_{i=1}^{I} \prod_{k=1}^{K} F(a_k \theta_i - b_k)^{y_{ik}} [1 - F(a_k \theta_i - b_k)]^{(1-y_{ik})}.$$  

(2.20)

This likelihood function has a total of $I+2K$ parameters, assuming the parametric form of $F$ is known.
2.4.2.2 Prior Distributions

As mentioned, we assign prior distributions to the parameters to model any prior beliefs about the parameters. Also, in IRT models in particular, the choice of prior distribution can help take care of issues with identifiability. For example, suppose we have ability parameter $\theta_i$ and discrimination parameter $a_k$ and another pair of parameters $\theta^*_i = 2\theta_i$ and $a^*_k = \frac{1}{2} a_k$. Then, in a 2PN case, we see that the probability of success for examinee $i$ on item $k$ is $p_{ik} = \Phi(a_k\theta_i - b_k) = \Phi(a^*_k\theta^*_i - b_k)$. This means that the two sets of parameters are not identifiable.

One solution to this identification problem is to place some constraints on the parameters. One such constraint is placed on the ability parameters, which is $\sum_{i=1}^{I} \theta_i = 0$. Another is on the item parameters which is that $\prod_{k=1}^{K} a_k = 1$ and $\sum_{i=1}^{I} b_k = 0$. Aside from this, the model also becomes identified by placing a known prior distribution on the ability parameters. Specifically, we place a $N(0,1)$ prior distribution on the ability parameters, as discussed in the beginning of section 2.3.

As mentioned, one of the modeling assumptions of IRT models is that the probability of correctly answering an item increases with ability level $\theta$. Thus, the discrimination parameters are assumed to be positive. Looking again at figure 2.2 we see that if an item has a very low discrimination parameter value then it is not very effective in discrimination between examinees of different ability. Likewise, very high discrimination parameter values mean an item is too precise and is also not useful. So, from a practical standpoint we want to restrict their values somewhat. We take discrimination parameter values between, say $1/2$ and $2$ as being useful (Johnson and Albert 1999, p.192).

Considering the item difficulty parameters, since we have placed a $N(0,1)$ prior distribution on the abilities it makes sense to assign a prior on the difficulty parameter $b_k$ which has small probability of being outside the range $(-3,3)$. Looking at figure 2.3 we can see that if $b_k$ was outside $(-3,3)$ then the item response curve would be shifted far to the left.
or right. This would mean that the probability of success would be approximately 1 or 0 for examinees with abilities within the range of \((-3, 3)\). That would mean having items that would almost always produce all correct or incorrect responses.

What we want is to specify vague prior distributions on the item parameters that will provide for these considerations. For the item discrimination parameters we assume a truncated normal prior distribution with mean 1 and variance 0.50, denoted \(TN_{\theta,\infty}(1, .50)\), where in general \(TN_{\theta,b}(\mu, \sigma^2)\) represents the \(N(\mu, \sigma^2)\) distribution truncated to the interval \((\theta, b)\). The truncation is done to the right of zero so we are \textit{a priori} restricting the discrimination parameters to be positive, and the mean and variance ensures a low probability of having values too large or small. For the item difficulty parameters we use a \(N(0,1)\) prior distribution which will place a high prior probability on the \(b_k\) being between \(-3\) and \(3\). These prior distributions are relatively weak statements about the location of the parameters.

### 2.4.2.3 Posterior Distribution of \(\theta, a,\) and \(b\)

Given the likelihood in (2.20), then the joint posterior distribution of the ability and item parameters can be written as

\[
\pi(\theta, a, b | y) \propto \prod_{i=1}^{I} \prod_{k=1}^{K} F(a_k \theta_i - b_k) y_{ik} [1 - F(a_k \theta_i - b_k)]^{(1-y_{ik})} \phi(\theta, a, b),
\]

(2.21)

where \(\phi(\cdot; \mu, \sigma^2)\) here represents the normal \(pdf\) with mean \(\mu\) and variance \(\sigma^2\) and \(g(\cdot; \mu, \sigma^2, l, u)\) represents the truncated normal \(pdf\) with mean \(\mu\), variance \(\sigma^2\), and lower and upper bounds.
of the interval of truncation $l$ and $u$.

We now break from our general description of the two-parameter Bayesian IRT model to discuss the Bayesian 2PN IRT model in more detail. In particular, this is the model that our research will deal with later in chapters 4 and 5.

### 2.4.3 Bayesian Two-Parameter Normal Ogive IRT Model

Here we discuss a little more specifically about the 2PN version of the two-parameter Bayesian IRT model. So, in this the link function $F$ used throughout section 2.4.2 will be set equal to the probit link function $\Phi(.)$. In this model, we will incorporate a data augmentation scheme which we will describe next.

#### 2.4.3.1 Binary Data Augmentation

In Albert (1992), a data augmentation scheme was introduced for the 2PN model in which an independent random latent variable $Z_{ik}$ can be incorporated into the data generating level of the model. Corresponding to each binary response $y_{ik}$, we assume there exists a continuous latent variable $Z_{ik}$ that determines the performance of the student on item $k$. In this, the relationship the variable $Z$ has with the data is that student $i$ answers item $k$ correctly if $Z_{ik} > 0$ and incorrectly if $Z_{ik} \leq 0$. The observed responses $y_{ik}$ can then be thought of as indicator variables for the $Z_{ik}$, where $y_{ik} = 1$ if $Z_{ik} > 0$, and $y_{ik} = 0$ if $Z_{ik} \leq 0$.

Albert (1992) introduced this latent variable $Z$ into the 2PN model where $Z_{ik}$ is considered to be distributed as a $N(a_k \theta_i - b_k, 1)$. We can see that with this data augmentation step the probability of correct response of a student on the $k$th item is the same as (2.7).
For example, with \( Z_{ik} \sim N(a_k \theta_i - b_k, 1) \) then,

\[
P(y_{ik} = 1|\theta_i, a_k, b_k) = P[Z_{ik} > 0|(a_k \theta_i - b_k), 1]
= P[Z_{ik} - (a_k \theta_i - b_k) > -(a_k \theta_i - b_k)|0, 1]
= P[Z_{ik}^* > -(a_k \theta_i - b_k)|0, 1], \quad Z_{ik}^* \sim N(0, 1)
= P(Z_{ik}^* < a_k \theta_i - b_k)
= \Phi(a_k \theta_i - b_k).
\]

There are a few reasons why we choose to include the latent variable \( Z \) in the model. One is that it allows for an easy implementation of a Gibbs MCMC sampling algorithm to simulate draws from the joint posterior distribution of \( \theta, a, b \), and \( Z \). In Gibbs sampling, the full conditional distributions of each variable must be known. Incorporating the latent variable \( Z \) in the model allows us to obtain the conditional distributions of each variable and parameter in closed form. We will discuss this Gibbs sampling algorithm in more detail in section 2.5.2.

Another reason for incorporating \( Z \) is that under a 2PN IRT model, with \( Z_{ik} \) defined as above with prior distribution \( N(a_k \theta_i - b_k, 1) \), then the conditional posterior distribution of \( Z_{ik} \) given \( \theta, a, \) and \( b \) is a truncated normal distribution. We can use this to define residuals based on \( Z \) to perform model checking. This will be discussed more in chapter 4.

### 2.4.3.2 Bayesian 2PN IRT Model Specification

We now give a specific form of the two-parameter Bayesian normal ogive IRT model. We include a data augmentation step as defined above with latent variable \( Z \), where \( Z_{ik} \) is
normally distributed with mean $a_k \theta_i - b_k$ and variance one. The model is specified here as,

$$y_{ik} = \begin{cases} 
1, & \text{if } Z_{ik} > 0 \\
0, & \text{if } Z_{ik} \leq 0 
\end{cases} \quad (2.23)$$

$$Z_{ik} \sim N(a_k \theta_i - b_k, 1) \quad (2.24)$$

$$\theta_i \sim N(0, 1) \quad (2.25)$$

$$a_k \sim T N(0, \infty) (1, .50) \quad (2.26)$$

$$b_k \sim N(0, 1), \quad (2.27)$$

for $i = 1, \ldots, I$ and $k = 1, \ldots, K$.

In the next section we will look at how parameter estimation is done under a frequentist and Bayesian approach. In this we will show a Newton-Raphson procedure for obtaining the joint maximum likelihood estimates of the ability and item parameters. Also, we will outline a Gibbs MCMC procedure mentioned at the end of section 2.4.3.1 for simulating samples of parameter values from their marginal posterior distributions. From here, point estimates can easily be obtained by computing sample means.

### 2.5 Parameter Estimation under Frequentist and Bayesian Approaches

One of the interests in statistical modeling is to estimate the parameters of the model given the observed data. In IRT models, we are interested in estimating the item and ability parameters. The Bayesian and frequentist methods have different approaches to this. The frequentist approach looks at maximizing the likelihood function. In the Bayesian approach inference about a parameter is based on the marginal posterior distribution. Via simulation, we can draw a representative sample from the marginal posterior distribution, and the sample mean can be computed as a point estimate of the parameter.
We will look at an example here using both methods to estimate the item and ability parameters of the 2PN IRT model. We will simulate a response data matrix under the 2PN model and compute a set of parameter estimates under each method. Using scatter plots, we will plot the true parameter values against the estimates to see how close the estimates are to the true values. At the end of the section we will then compare the performance of the frequentist and Bayesian estimates and discuss some advantages of the Bayesian approach. The frequentist approach will be discussed first, followed by the Bayesian method.

2.5.1 Frequentist Approach

As mentioned, the frequentist approach uses maxima of the likelihood functions as parameter estimates. The reasoning for this is that the likelihood function is a function of the model parameters which gives the likelihood of observing data $y$. Since we are interested in estimating the values of the model parameters that produced the data $y$, it seems natural to use the values that maximize the likelihood function, i.e. the values that give the greatest likelihood of producing $y$. We derived the likelihood function for the two-parameter model in (2.20).

In maximizing the likelihood function it is algebraically easier to maximize the natural log of the likelihood (log-likelihood). Taking the natural log of both sides of (2.20) we get

$$\ln L = \sum_{i=1}^I \sum_{k=1}^K \left[ y_{ik} \ln P_{ik} + (1 - y_{ik}) \ln(1 - P_{ik}) \right],$$

(2.28)

where $P_{ik} = \Phi(a_k \theta_i - b_k)$. The values that maximize this function are found by taking the first-order partial derivatives with respect to each parameter and solving the equations,

$$\frac{\partial \ln L}{\partial a_k} = 0, \quad \frac{\partial \ln L}{\partial b_k} = 0, \quad \frac{\partial \ln L}{\partial \theta_i} = 0.$$  

(2.29)

A full derivation of these equations for the 2PN (and 2PL) IRT model can be found in Baker (2004). It is also discussed that these equations can not be solved analytically and so must
be done using numerical methods. Two methods commonly used are the Newton-Raphson procedure and Fisher’s Scoring Method.

Under the Newton-Raphson procedure, for a function $f(x)$ to be maximized, an initial value $x_0$ is given and based on this value the Newton-Raphson equation is used to find an improved estimate $x_1$, and so on. The Newton-Raphson equation is given by,

$$x_{i+1} = x_i - [f'(x_i)]^t [f''(x_i)]^{-1} i = 0, 1, 2, \ldots, \tag{2.30}$$

where $f'(\cdot)$ is the gradient vector of first-order partial derivatives and $f''(\cdot)$ is the hessian matrix of second-order partial derivatives. As mentioned, from initial estimate $x_0$ we find an improved estimate $x_1$, and the process is repeated until the difference between successive estimates is sufficiently close to zero.

Fisher’s Scoring Method is an alternative method that is performed in the same manner as the Newton-Raphson procedure. The only difference is that the negative of the hessian matrix in (2.30) is replaced by the expected value of the hessian matrix. This is usually referred to then as the Fisher’s equation. The advantage of using the Fisher’s equation versus the Newton-Raphson equation is that it may converge to a solution faster (Baker 2004, p.34).

Birnbaum (1968) introduced a joint maximum likelihood estimation (JMLE) procedure to estimating item and ability parameters of two-parameter IRT models. In this procedure, the item and ability parameters are given initial estimates based on the observed data. Then, starting with the item parameters, the examinee abilities are considered known and a Newton-Raphson procedure, or Fisher’s Scoring Method is used to get maximum likelihood estimates (MLEs) of item parameters $a$ and $b$. Next, these item parameter estimates are treated as known and the same procedure is used to find the MLEs of the ability parameters $\theta$. This process is repeated iteratively until some convergence criteria among all parameters is met, typically when the maximum absolute difference between estimates from one iteration
to the next becomes small.

We will now show an example illustrating this procedure where we compute the JMLEs for the true parameter values of response data generated under a 2PN IRT model. We generated a response data matrix \( Y \) of \( I = 200 \) examinees to an examination with \( K = 40 \) items from the 2PN model described in section 2.4.3.2. The data generating item discrimination and difficulty parameters were simulated according to (2.26) and (2.27), respectively. The data generating ability parameters were simulated from a standard normal distribution according to (2.25). After simulating parameters \( \theta_i, i = 1, \ldots, I, a_k \) and \( b_k, k = 1, \ldots, K \), we simulate latent variables \( Z_{ik} \sim N(a_k \theta_i - b_k) \) and finally generate the data \( y_{ik} \) where \( y_{ik} = 1 \) if \( Z_{ik} > 0 \) and \( y_{ik} = 0 \) if \( Z_{ik} \leq 0 \).

To find the JMLEs \( \hat{\xi}^* \) and \( \hat{\theta}^* \) of the item and ability parameters we use Fisher’s Scoring Method. A complete explanation of this procedure for a 2PN IRT model along with full derivations of the Fisher’s equations used can be found in Baker (2004).

To compare the JMLEs \( \hat{\xi}^* \) and \( \hat{\theta}^* \) and parameter values, we create scatter plots given in figures 2.5 and 2.6. Figure 2.5 shows scatter plots between the true item parameters and their corresponding JMLEs. The left side of figure 2.5 plots the true discrimination parameters against their corresponding JMLEs, and the right side plots the true difficulty parameters against their JMLEs. The line \( y = x \) is plotted in each for reference.

In the left plot of figure 2.5 we see the estimates seem fairly close to the true parameter values for small values of the item discrimination \( a \). However, a discrepancy is apparent for larger values of \( a \) where there is a noticeable positive bias in the estimates. As an overall measure of how close the JMLEs are to the true values we compute the correlation coefficient between them, which is 0.8922. The closer this number is to one then the closer the plotted points are to following a perfect linear trend, meaning the closer their respective coordinate values are to being equal. Also, in the next section we will be looking at Bayesian posterior estimates of these parameters and we can compare the correlation coefficients to see which set of estimates is closer to the true values overall.
Figure 2.5 Scatter plots of the 40 true item discrimination parameters (left) and item difficulty parameters (right) and their corresponding joint maximum likelihood estimates.

In the right plot of figure 2.5 we see that there is a linear trend of points around the line $y = x$. The correlation coefficient between the difficulty JMLEs and the true values is computed to be 0.8612. From this, the JMLEs seem to do a decent job of estimating the true difficulty parameters. The exception is for a few points which lie far above the line $y = x$, indicating the estimates are much greater than the true values. This actually highlights a problem with item difficulty JMLEs in IRT. If many examinees got an item incorrect, the JMLE of the difficulty will tend to overestimate how difficult the item is. For example, item 20 has a JMLE close to 3 where the true value is close to 1.2. This item only had 13% correct responses. An extreme case of this is when an item is scored incorrect by all examinees in which case the estimate would be positive infinity. A similar case holds for items that receive most to all correct responses.

Figure 2.6 shows the scatter plots of the 200 ability JMLEs versus the true values. We see a nice linear pattern around the line $y = x$. The correlation coefficient between the ability JMLEs and the true values is computed to be 0.9466. The JMLE’s seem to underestimate the larger true ability values. Having larger variability in estimates of extreme values sometimes occurs with the maximum likelihood estimates.
2.5.2 Bayesian Approach

As mentioned, inference about a parameter under the Bayesian approach is done through the looking at the marginal posterior distribution. We specified the joint posterior distribution of the ability and item parameters under the 2PN IRT model in (2.22). This distribution has an intractable form, but a common alternative is to use simulation to generate samples from the joint posterior distribution. From this, we can summarize the simulated draws for each parameter to make inferences about that parameter. The sample mean can be computed as an estimate of the marginal posterior mean, which is also used as a point estimate of the parameter. Computing approximate 95% probability intervals on the posterior samples can also be used to do inference about the parameters, but we are only looking at point estimates for now.

An effective means of simulating from a multivariate distribution is known as the Gibbs sampling algorithm. This is a type of a Markov chain Monte Carlo (MCMC) procedure, where to simulate a sample from say, \( h(x, y, z) \) (the joint distribution of random variables \( X, Y, \) and \( Z \)) draws are iteratively taken from the full conditional distributions of each variable. More specifically, for full conditional distributions \( w(x|y, z), w(y|x, z), \) and \( w(z|x, y) \) we start...
with initial values $x_0, y_0,$ and $z_0$ and draw in order,

\begin{align*}
x_1 & \sim w(x|y_0, z_0) \\
y_1 & \sim w(y|x_1, z_0) \\
z_1 & \sim w(z|x_1, y_1),
\end{align*}

(2.31) (2.32) (2.33)

to give draw $(x_1, y_1, z_1)$. Note that the values drawn from one conditional distribution are updated in the next conditional distribution before simulating the next variable. This gives us one iteration of the Gibbs sampling algorithm. This draw is now treated as the values $x_0, y_0,$ and $z_0$ and the above steps are repeated for further iterations.

We should note that the draws taken when first starting the algorithm are not necessarily distributed under the desired distribution $h(x, y, z)$. Generally, the process is repeated for say, $m$ iterations until the draws converge to the target distribution $h(x, y, z)$. This is called the “burn in” period. After this point the algorithm is continued for another $L$ iterations to obtain a sample with distribution $h(x, y, z)$. Now, from the theory of Markov chains this sample of size $L$ is dependent. A common practice used to address this is to “thin” the after burn in sample. This is done by specifying some positive integer $n$, and then only keeping every $n^{th}$ draw until the sample of size $L$ is reached. This process helps to reduce the dependence among the overall sample values.

In order to use Gibbs sampling, one needs to know the full conditional distributions of each random variable involved. In our 2PN model, this translates to knowing the full conditional posterior of each parameter. As mentioned in section 2.4.3.1 Albert (1992) showed that with incorporating the latent variable $Z$ into the model, the full conditional posterior distributions of each variable and parameter can be specified in closed form. If we include the data augmentation step as we do in section 2.4.3.2 the joint posterior density is
\[
\pi(\theta, Z, a, b|y) \propto \prod_{i=1}^I \phi(\theta_i; 0, 1) \prod_{k=1}^K [\phi(Z_{ik}; m_{ik}, 1)(I(Z_{ik} > 0)I(y_{ik} = 1)
\quad + I(Z_{ik} \leq 0)I(y_{ik} = 0))] \prod_{k=1}^K g(a_k; 1, 50, 0, \infty)\phi(b_k; 0, 1).
\]

Specifically, we can specify the conditional posterior distributions: \( w(Z_{ik}|\theta, \xi, y) \), \( w(\theta_i|Z, \xi, y) \), and \( w(\xi_k|Z, \theta, y) \) for \( i = 1, \ldots, I \) and \( k = 1, \ldots, K \) and \( \xi = (a, b)^t \). The conditional posterior of \( Z_{ik} \), given \( (\theta, \xi, y) \), is truncated normal with mean \( m_{ik} = a_k \theta_i - b_k \) and variance 1. The truncation is from the left of 0 if \( y_{ik} = 1 \), and from the right of 0 if \( y_{ik} = 0 \). More on the conditional posterior distribution of \( Z \) will be discussed in chapter 3.

The conditional distribution of \( \theta_i \), given \( (Z, \xi, y) \), is a normal distribution with mean and variance given by,

\[
\mu_{\theta_i} = \frac{\sum_{k=1}^K a_k (Z_{ik} + b_k)}{\sum_{k=1}^K a_k^2 + 1}, \quad \text{and} \quad \sigma_{\theta_i}^2 = \frac{1}{\sum_{k=1}^K a_k^2 + 1}. \tag{2.35}
\]

The conditional posterior distribution of \( \xi_k = (a_k, b_k)^t \), given \( (Z, \theta, y) \), is bivariate normal with mean,

\[
\mu_{\xi_k} = \left[ X^t X + \Sigma_0^{-1} \right]^{-1} \left[ X^t Z_k + \Sigma_0^{-1} \mu_\xi \right], \tag{2.36}
\]

and covariance matrix

\[
\Sigma_{\xi_k} = \left( X^t X + \Sigma_0^{-1} \right)^{-1}, \tag{2.37}
\]

where

\[
\mu_\xi = (\mu_a, \mu_b)^t \quad \text{and} \quad \Sigma_0 = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}, \tag{2.38}
\]

with \( \mu_a, \sigma_a^2 \) and \( \mu_b, \sigma_b^2 \) being the prior means and variances of \( a \) and \( b \), respectively, \( X = [\theta - 1] \) and \( Z_k = (Z_{1k}, \ldots, Z_{Ik})^t \).

Putting this all together, we implement the Gibbs sampling algorithm by first choosing initial values \( Z_{ik}^{(0)}, \theta_i^{(0)} \), and \( \xi_k^{(0)} = (a_k^{(0)}, b_k^{(0)})^t \) for \( i = 1, \ldots, I \) and \( k = 1, \ldots, K \) (using the
prior means is usually sufficient). Then for \( m = 0, 1, 2 \ldots \),

1. 

\[
Z_{ik}^{(m+1)}|\theta^{(m)}, \xi^{(m)}, y \sim \begin{cases} 
TN_{(0,\infty)}(m_{ik}, 1), & \text{if } y_{ik} = 1 \\
TN_{(-\infty,0)}(m_{ik}, 1), & \text{if } y_{ik} = 0 
\end{cases}
\]  

(2.39)

2. 

\[
\theta_i^{(m+1)}|Z^{(m+1)}, \xi^{(m)}, y \sim N(\mu_{\theta_i}, \sigma_{\theta_i}^2)
\]  

(2.40)

3. 

\[
\xi_k^{(m+1)}|Z^{(m+1)}, \theta^{(m+1)}, y \sim N(\mu_{\xi_k}, \Sigma_{\xi_k}).
\]  

(2.41)

We use this algorithm to simulate from the joint posterior in (2.34) given the generated “observed” data \( Y \) from the previous section. We simulate a total of 7,000 posterior draws. A burn in length of 4,000 is used and the remaining 3,000 draws are thinned by 3 to give an end posterior sample of 1,000 draws. Again, this means that we have a sample of 1,000 draws for each parameter from its marginal posterior. We compute the mean of each sample and use these as estimates of the true parameter values used to generate the data.

To see how close these posterior parameter estimates are to the true values we again create scatter plots given in figures 2.7 and 2.8. The left plot in figure 2.7 shows the posterior estimates of the item discrimination parameters versus the true values, and the right plot shows the posterior estimates of the difficulty parameters versus the true values. A linear trend is followed along the line \( y = x \) in both plots, with the trend being very strong for the right plot. We compute the correlation coefficients between the true discrimination and difficulty parameters and their corresponding posterior estimates to be 0.9419 and 0.9906, respectively. This shows how strong the linear trend is in each plot.

In figure 2.8 we also see a nice linear trend around the line \( y = x \). The correlation coefficient is computed as 0.9706. There is a noticeable improvement in these estimates over
the ability JMLEs for larger values of $\theta$. In the next section we will compare the frequentist and Bayesian estimates we have obtained so far, but for now we can see that the plotted points lie closer to the line $y = x$ in figure 2.8 than in figure 2.6 for larger true values of $\theta$.

Figure 2.7 Scatter plots of the 40 true item discrimination parameters (left) and item difficulty parameters (right) and their corresponding posterior estimates.

Figure 2.8 Scatter plot of the 200 true ability parameters and their posterior estimates.
2.5.3 Comparing JMLE and Bayesian Parameter Estimates

2.5.3.1 Similarity of MLEs and Posterior Means

We first might ask why we are able to compare maximum likelihood estimates with posterior means? That is, why would we expect a priori that the posterior means and MLEs should be similar? Recall from (2.16) that the posterior distribution \( \pi(\omega|y) \) of parameter \( \omega \) given data \( y \) is proportional to the likelihood times the prior distribution of \( \omega \). Another way to view the posterior is the likelihood weighted by the prior density. If the prior weight is very diffuse, then it will have very little effect on the likelihood. This means that the likelihood should not be influenced much by the prior distribution and so the posterior distribution and likelihood should be similar. Also considering the data sample size \( n \), Walker (1969) shows that as \( n \) tends to infinity the mean of the posterior distribution \( \pi(\omega|y) \) approaches the MLE of \( \omega \).

We can show this with an example where we consider the model \( y_1, \ldots, y_n \sim N(\omega, \sigma^2) \) with prior distribution \( \omega \sim N(\mu, \tau^2) \), with \( \sigma^2 \), \( \mu \), and \( \tau^2 \) known. The posterior distribution of \( \omega \) can be shown to be a normal distribution with mean and variance

\[
\tilde{\mu} = \frac{\tau^2}{\sigma^2/n + \tau^2} \bar{y} + \frac{\sigma^2/n}{\sigma^2/n + \tau^2} \mu \quad \text{and} \quad \tilde{\tau}^2 = \frac{\tau^2 \sigma^2/n}{\sigma^2/n + \tau^2}. \tag{2.42}
\]

We can see that as we take larger values for the prior variance, then the posterior mean \( \tilde{\mu} \) shrinks towards \( \bar{y} \), which is the MLE of \( \omega \). We illustrate this by plotting the prior density, likelihood, and posterior density in figure 2.9, where we take \( \sigma^2 = 5 \), \( \mu = 0 \) and let \( \tau^2 \) be 1, 5, and 10. The plots in figure 2.9 are given from left to right for each value of \( \tau^2 \), respectively. We see that as the prior variance \( \tau^2 \) increases, the prior density becomes flatter and the posterior density approaches the likelihood.

Consider also the sample size \( n \) involved in \( \tilde{\mu} \) in (2.42). If the sample size \( n \) is allowed to tend toward infinity then we also get that the overall value shrinks towards the MLE \( \bar{y} \). We
Figure 2.9 Plots of the prior density, likelihood, and posterior density for the model $y_1, \ldots, y_n \sim N(\omega, \sigma^2)$ with prior distribution $\omega \sim N(\mu, \tau^2)$. We take $\sigma^2 = 5$, $\mu = 0$, and take $\tau^2 = 1$ (left), $\tau^2 = 5$ (center), and $\tau^2 = 10$ (right). As the prior variance $\tau^2$ increases, the prior density flattens and the posterior approaches the likelihood.

illustrate this by plotting the prior density, likelihood, and posterior density in figure 2.10 where we take $\sigma^2 = 5$, $\mu = 0$ and $\tau^2 = 1$ and consider sample sizes of $n = 10, 30, \text{ and } 50$. The plots in figure 2.10 are given from left to right for each value of $n$, respectively. We see that as $n$ increases, the posterior density approaches the likelihood.

Figure 2.10 Plots of the prior density, likelihood, and posterior density for the model $y_1, \ldots, y_n \sim N(\omega, \sigma^2)$ with prior distribution $\omega \sim N(\mu, \tau^2)$. We take $\sigma^2 = 5$, $\mu = 0$, and $\tau^2 = 1$. We consider three sample sizes of $n = 10$ (left), $n = 30$ (center), and $n = 50$ (right). As the sample size $n$ increases, the posterior approaches the likelihood.
2.5.3.2 Comparison of JML and Bayesian Item and Ability Parameter Estimates

We compare the JMLEs and Bayesian parameter estimates by seeing how close the respective sets of estimates are to the true parameter values. We can recall from sections 2.5.1 and 2.5.2 that when plotting the scatter plots of the JMLEs in figures 2.5 and 2.6 and Bayesian estimates in figures 2.7 and 2.8 against the true values of the parameters, we also calculated the correlation coefficients. Here we compare these for the discrimination, difficulty, and ability parameters in table 2.1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>JMLEs</th>
<th>Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>0.8922</td>
<td>0.9419</td>
</tr>
<tr>
<td>$b$</td>
<td>0.8612</td>
<td>0.9906</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.9466</td>
<td>0.9706</td>
</tr>
</tbody>
</table>

Table 2.1 Computed correlation coefficients between the set of true discrimination parameters $a$, difficulty parameters $b$, and ability parameters $\theta$, and their respective sets of joint maximum likelihood and Bayesian estimates.

The first column in table 2.1 gives the correlation coefficients between the JMLEs and true parameter values, and the second column gives the correlation coefficients between the Bayesian posterior estimates and true parameter values. We see that for all sets of parameters, the correlation coefficients calculated between the Bayesian estimates and true values are higher than those for the JMLEs. This implies that the linear trend in the scatter plots dealing with the Bayesian estimates was stronger than in those corresponding to the JMLEs. This in turn means that overall, the Bayesian estimates seem to be closer to the true parameter values than the JMLEs.
Figure 2.11 Difference plot for item discrimination parameter estimates. The difference between the estimates and true values (estimate - true value) is plotted against the true parameter values.

To look at this closer, we create difference plots in figures 2.11, 2.12, and 2.13 where the difference between the estimates and true values (estimate - true value) is plotted against the true parameter values. The left side in all three figures gives the difference plot for the JMLEs and the right side gives the plots for the Bayesian estimates. In all plots the horizontal line $y = 0$ is plotted for reference.

In figure 2.11 we see that most of the points appear to lie from $-0.50$ to $0.50$ on the horizontal scale in both plots. This indicates that the discrimination parameter estimates under both methods perform about the same as far as how close they are to the true values. However, there are some discrimination parameters with JMLEs that severely overestimate the true values, and this does not happen with the Bayesian estimates.

In figure 2.12 we see that overall the points in the right plot lie much closer to the $y = 0$ line than in the right plot. This shows that the Bayesian estimates are much closer to the true values than the JMLEs. Following from this, we note that the items with extreme JMLE difficulty estimates (discussed in section 2.5.1) have much better Bayesian estimates.

In figure 2.13 we see that most of the points seem to lie between $-0.70$ and $0.70$ on the horizontal scale in both plots. The improvement in the Bayesian estimates is seen for those
Figure 2.12 Difference plot for item difficulty parameter estimates. The difference between the estimates and true values ($estimate - true value$) is plotted against the true parameter values.

true abilities that are large or small. As also discussed in section [2.5.1](#2.5.1), there can be more variability in maximum likelihood estimates of large or small ability values. This can be seen by the few points corresponding to small or large ability values in the left plot. The points in the right plot corresponding to the Bayesian estimates of these values are much closer to the line $y = 0$. This indicates these estimates are closer to the true values.

### 2.5.4 Advantages of the Bayesian Approach

As we can see, there are advantages of the Bayesian method versus the JMLE method. As mentioned, JMLEs of item parameters can be extreme when an item is answered correctly or incorrectly by most or all examinees. An item in which all examinees answer correctly or incorrectly can have a difficulty JMLE of positive or negative infinity, respectively. Likewise, an examinee who answered all items correct or incorrect would get a JMLE ability estimate of positive infinity or negative infinity, respectively. This occurrence does not happen with the Bayesian estimates due to the use of the prior distributions. This is because the prior distributions placed on the item and ability parameters help restrict parameter values to a reasonable range. For example, the $N(0, 1)$ prior placed on the ability and difficulty
Figure 2.13 Difference plot for item ability parameter estimates. The difference between the estimates and true values \((\text{estimate} - \text{true value})\) is plotted against the true parameter values.

parameters discourages values outside the range of \((-3, 3)\). Thus, under the aforementioned conditions, infinite estimates do not occur.

As mentioned, identifiability is an issue when working with a 2PN IRT model. To solve this, certain constraints can be placed on the parameters as mentioned in section 2.4.2.2. From a classical approach, this can be avoided by using the Marginal Maximum Likelihood (MML) method. This method uses the marginal likelihood function which is obtained by integrating out the joint likelihood function with respect to all the ability parameters. The only problem is that this method does not allow one to obtain ability estimates. The Bayesian method however, allows the estimation of all parameters while the identifiability problem is solved by use of the known standard normal prior distribution on the ability parameters.

There are some other advantages which we will mention here, but will expound on further in subsequent chapters. One is from a standpoint of inference. In frequentist statistics, the computation of say a 95% confidence interval of a parameter of interest is a common tool. The interpretation of this can be somewhat awkward. That is, if the experiment were repeated many times with multiple samples of data being obtained and an interval estimate is computed for each, over the long run the true parameter value would lie in this interval.
approximately 95% of the time. This is less clear than a Bayesian probability interval which can be computed from the posterior distribution. Since we can directly simulate from the posterior distribution of a parameter, we can make a direct probability statement about what values the parameter lies between. For example, based on a simulated posterior sample that has a 0.025 quantile equal to $-2$ and 0.975 quantile equal to 2, we can say there is an approximate 95% probability that the parameter is between $-2$ and 2. This makes interpretation much easier.

We have another advantage with a Bayesian approach when doing model checking, which will be the focus of our work in chapters 4 and 5. In a frequentist approach, model checking is often done by computing the value of a statistic based on a set of observed data and comparing this to some reference distribution. This reference distribution is often the sampling distribution of the statistic, if it is known. If it is not known, then an asymptotic approximation to this distribution must be used. In chapter 4 we will look at many classical statistics used for checking IRT model fit. A problem we will see is that the exact sampling distributions of these are unknown, making any inference done with them in a classical setting questionable. Under a Bayesian approach, one can make use of predictive distributions, namely the prior predictive and posterior predictive distributions (to be introduced in chapter 3). Using simulation, we can directly draw samples from these distributions to perform inference with using the observed value of the statistic.

In the next chapter we will introduce some methods for performing model checking with Bayesian models. In particular, we will look at two standard methods involving the predictive distributions mentioned above. Then we will introduce two newer methods that can be used in Bayesian model checking. The discussion in the next chapter will be a general one, meant to give the reader an idea of what the different methods are and how they are employed in checking aspects of fit in Bayesian models. Later, in chapters 4 and 5 we will discuss our research on how these methods can be used with Bayesian IRT models.
CHAPTER 3

BAYESIAN MODEL CHECKING METHODS

3.1 Introduction

In this chapter we will discuss several methods for performing model checking for Bayesian models. Model checking is different than model selection. When performing model checking, we are looking to check, or assess, the fit of a particular model to a given set of data. In model selection, we are looking to choose a best fitting model.

Model checking is performed in both a frequentist and Bayesian setting. In this, for data $Y$ we have a proposed model $M$. Under a frequentist model $M = M_F$ we could have $Y \sim f(y|\theta)$ where $f(y|\theta)$ is the sampling density of $Y$. Under a Bayesian model $M = M_B$ we could have $Y \sim f(y|\theta) ; \theta \sim p(\theta)$ where $f(y|\theta)$ is the sampling density of $Y$ and $p(\theta)$ is the prior distribution of the parameter $\theta$. For checking the fit of a model, in either case, we suppose a null hypothesis that the data is distributed according to the model, but have no specific alternative.

In this chapter we will focus on model checking methods for Bayesian models. We will start by discussing two common Bayesian model checking methods based on predictive distributions. We will then introduce and discuss two newer Bayesian methods that can be used for model checking. The goal of this chapter will be to give the reader a general understanding of the different methods and how they can be used for checking aspects of fit in Bayesian models. Chapters 4 and 5 will discuss our research where we apply some of these methods to checking aspects of fit in Bayesian IRT models.
### 3.2 Bayesian Predictive Model Checking

First consider a frequentist setting where a statistical model $H_0 : Y \sim f(y|\theta)$ is being considered. Given observed data $y_{\text{obs}}$ we are interested in checking whether the model is adequate. The common approach is to choose a test statistic $T(Y)$ where, without loss of generality, large values of $T$ evaluated at $y_{\text{obs}}$ indicate inadequacy of the model in describing $y_{\text{obs}}$. A common way of assessing the extremeness of the value of $T(y_{\text{obs}})$ is to compute the tail area probability, or $p$-value, of $T$ exceeding the value of $T(y_{\text{obs}})$. This $p$-value is given as,

$$p(y_{\text{obs}}) = P_T(T(Y) \geq T(y_{\text{obs}})|\theta),$$  \hspace{0.5cm} (3.1)

where the probability computation is done with respect to $P_T$, the reference distribution of $T$. In a classical setting the reference distribution for $T$ is its sampling distribution, given data under the model.

Now, we are not going to discuss frequentist model checking in depth, but discuss it here to motivate using the two Bayesian methods we will introduce next. However, before going on to these we will mention one difficulty with the $p$-value in (3.1), and this has to do with the case when $\theta$ is unknown, i.e. when there are nuisance parameters. One nice situation where this is taken care of is if $T$ is ancillary. That is, the distribution of $T(Y)$ is free of $\theta$. If $T$ is not ancillary, then computation of the $p$-value requires some way of eliminating $\theta$ from (3.1). Some of the ways of doing this are reviewed in Bayarri and Berger (2000) and Robins et al. (2000).

We move on to discuss two commonly used Bayesian model checking methods: the prior predictive method and posterior predictive method. Now, we are considering a Bayesian model $H_0 : Y \sim f(y|\theta); \theta \sim p(\theta)$ to fit to the data $y_{\text{obs}}$. We will see in these methods how the problem of dealing with nuisance parameters in (3.1) is solved.

As in frequentist statistics, we start by choosing some discrepancy measure (Gelman et al. 1996) $T$, for which large or small values indicate a discrepancy between the model and data...
We note here that in a Bayesian setting we will use the term discrepancy measure instead of calling $T$ a test statistic. This is to avoid confusion on how $T$ can be computed. The strict definition of a statistic is that it is a function of the data. Here, a discrepancy measure can be a function of the data only, or data and model parameter(s). To assess the extremeness of the value of $T$ evaluated given $y_{obs}$, we can use its prior predictive or posterior predictive distribution as a reference distribution.

### 3.2.1 Prior Predictive Assessment

Box (1980) proposed using the prior predictive distribution for use in model checking. The prior predictive distribution (also, marginal distribution, or marginal predictive distribution) of $y$ is described by the density

$$f(y) = \int f(y|\theta)p(\theta)d\theta,$$

where $f(y|\theta)$ is the sampling density for $y$, and $p(\theta)$ is the prior density of $\theta$. We note here that we are integrating out $\theta$ in [3.2], so the prior predictive density $f(y)$ does not depend on $\theta$. Also, it is of importance to note that if data $y$ is distributed under the prior predictive distribution, then $T(y)$ is distributed under the prior predictive distribution of $T$.

As mentioned, the prior predictive distribution of $T$ is used as a reference distribution for which to compare the value of $T(y_{obs})$. As in the classical setting we compute a tail area probability, or $p$-value, given here as

$$p_{prior}(y_{obs}) = P(T(y) \geq T(y_{obs})),$$

where the probability computation is done with respect to the prior predictive distribution of $T$. We note here that since we have integrated out $\theta$ in [3.2], the $p$-value in [3.3] is free of $\theta$. This is an advantage of this Bayesian method (as well as the posterior predictive method) in that the issue of nuisance parameters is easily taken care of.
In practice, we can estimate the $p$-value in (3.3) by simulation. We can simulate a draw of $y$ from its prior predictive distribution by: (1) simulating parameter(s) $\theta$ from its prior distribution $p(\theta)$ and (2) simulate $y$ from its sampling density $f(y|\theta)$. We then can compute the value of $T(y)$ as a draw from its prior predictive distribution. If we repeat this process, say $R$ times, we generate a sample $T(y_1), \ldots, T(y_R)$ from the prior predictive distribution of $T$. We can use this to estimate the probability in (3.3) by looking at the proportion of simulated values of $T(y_r)$, $r = 1, \ldots, R$, that are greater than the observed value of $T(y_{obs})$.

It is important to note that the prior distribution $p(\theta)$ must be a proper probability distribution (i.e. integrates to one) to perform prior predictive assessment. The reason for this is that if $p(\theta)$ is improper (i.e. integrates to infinity) then the prior predictive distribution in (3.2) will not be proper and so the $p$-value in (3.3) can not be computed. We will expound a little here, but the reader may see Gelman (2002) for a brief exposition on the different types of prior distributions (informative, non-informative, etc.) in Bayesian models. In some cases when researchers may not have much faith in their prior beliefs, they may choose to use a vague (or diffuse) prior distribution that is improper. In this, they are letting the inference be driven by the data with not much influence from the prior specification. However, if $p(\theta)$ is improper then integrating the prior predictive distribution $f(y)$ with respect to $y$ gives,

$$\int f(y)dy = \int \left( \int f(y|\theta)p(\theta)d\theta \right) dy = \int p(\theta) \left( \int f(y|\theta)dy \right) d\theta = \int p(\theta)d\theta,$$  \hspace{1cm} (3.4)

which is infinite since $\int p(\theta)d\theta$ is infinite. The next Bayesian model checking method, the posterior predictive method, allows for use of improper priors.

### 3.2.2 Posterior Predictive Assessment

Guttman (1967) and Rubin (1984) are among some who introduced the idea of using the posterior predictive distribution for model checking. Here we use the posterior predictive distribution of the discrepancy measure $T$ as a reference distribution. The posterior predic-
tive distribution of $T$ is computed by first computing the posterior predictive distribution of $\mathbf{y}$. The distribution of the posterior predictive data, $\tilde{\mathbf{y}}$, is given by

$$f(\tilde{\mathbf{y}}|\mathbf{y}_{obs}) = \int f(\tilde{\mathbf{y}}|\theta)\pi(\theta|\mathbf{y}_{obs})d\theta,$$  \hspace{1cm} (3.5)$$

where we are averaging the sampling density of $\tilde{\mathbf{y}}$ over the posterior distribution of the parameters. Here we briefly continue with the comment made at the end of the last section [3.2.1] Improper prior distributions may be used, because when formally combined with the data likelihood it may yield an acceptable proper posterior distribution. This in turn yields a proper posterior predictive distribution, which means the $p$-value given below in (3.6) will always be defined.

A posterior predictive $p$-value can be computed as,

$$p_{\text{post}}(\mathbf{y}_{obs}) = P\left(T(\tilde{\mathbf{y}}) \geq T(\mathbf{y}_{obs})\right).$$  \hspace{1cm} (3.6)$$

We make note again that the integration in (3.5) is being done with respect to $\theta$. Hence, the posterior predictive distribution $f(\tilde{\mathbf{y}}|\mathbf{y}_{obs})$ does not depend on $\theta$ and neither does the $p$-value $p_{\text{post}}(\mathbf{y}_{obs})$.

In practice we can estimate $p_{\text{post}}(\mathbf{y}_{obs})$ by simulation. We can simulate a draw of posterior predictive data $\tilde{\mathbf{y}}$ by: (1) simulate parameter(s) $\theta$ from their posterior distribution and (2) simulate data $\tilde{\mathbf{y}}$ from the sampling density $f(\tilde{\mathbf{y}}|\theta)$. We then compute the value of $T(\tilde{\mathbf{y}})$ as a draw from its posterior predictive distribution. If we repeat this process, say $R$ times, we generate a sample $T(\tilde{\mathbf{y}}_1), \ldots, T(\tilde{\mathbf{y}}_R)$ from the posterior predictive distribution of $T$. We can use this to estimate the probability in (3.6) by looking at the proportion of simulated values of $T(\tilde{\mathbf{y}}_r), r = 1, \ldots, R$, that are greater than the observed value of $T(\mathbf{y}_{obs})$.

Gelman et al. (1996) discussed the use of a discrepancy measure that depends on the data and the parameters $\theta$. We denote this discrepancy measure as $D(\mathbf{y}, \theta)$ and has posterior
predictive distribution based on the joint distribution of \( \tilde{y} \) and \( \theta \),

\[
f(\tilde{y}, \theta | y_{obs}) = f(\tilde{y}| \theta) \pi(\theta | y_{obs})
\]  

(3.7)

We see that (3.5) is just the marginal distribution of (3.7). It should also be clarified here that \( \theta \) is distributed under the posterior distribution.

We can define a similar tail area probability to (3.6). A tail area probability, or \( p \)-value can be defined as:

\[
p_{post}(y_{obs}, \theta) = P_{D}[D(\tilde{y}, \theta) \geq D(y_{obs}, \theta) | y_{obs}],
\]  

(3.8)

where the probability is taken over the joint distribution of \( (\tilde{y}, \theta) \). We can estimate \( p_{post}(y_{obs}, \theta) \) in practice by simulation: We do this by: (1) for given observed data \( y_{obs} \) we compute the posterior of \( \theta \) and draw a sample \( \theta_1, \ldots, \theta_n \) and (2) for each \( \theta_i \), for \( i = 1, \ldots, n \), we draw a sample of replicated data \( \tilde{y}_i \) from the sampling density \( f(\tilde{y}_i | \theta_i) \) and (3) compute the values of the discrepancy measure \( D(\tilde{y}_i, \theta_i) \), and \( D(y_{obs}, \theta_i) \) for \( i = 1, \ldots, n \). We then compute the pair-wise proportion

\[
\frac{\sum_{i=1}^{n} I[D(\tilde{y}_i, \theta_i) \geq D(y_{obs}, \theta_i)]}{n}
\]  

(3.9)

where \( I(.) \) is an indicator function. As before, \( p \)-values near 0 or 1 are taken as evidence of lack-of-fit.

### 3.2.3 Example of Prior Predictive and Posterior Predictive \( p \)-values

Here we show a simple example where we compute the prior predictive and posterior predictive \( p \)-values to determine model fit to a set of data. We will start by generating an “observed” data set \( y_{obs} = (y_1, \ldots, y_{50})^t \) where \( y_i \sim \chi^2_{20} \). We then consider fitting the null model \( y_i \sim N(\theta, \sigma^2), \), \( i = 1, \ldots, 50 \) where \( \theta \) has prior distribution \( N(\mu_0, \tau_0^2) \), with \( \sigma^2, \mu_0, \) and \( \tau_0^2 \) known. The prior distribution for \( \theta \) is conjugate, meaning that the posterior distribution
of $\theta$, $\pi(\theta | y_{obs})$, is in the same family as the assumed sampling density for $y_{obs}$. So, the posterior distribution $\pi(\theta | y_{obs})$ will be normal where the mean and variance can be derived as,

$$
\tilde{\mu}_n = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}
$$

and

$$
\tilde{\tau}_n^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2},
$$

(3.10)

where $\bar{y}$ is the sample mean of $y_{obs}$ and $n$ is the data sample size. For our example we will take $\sigma^2 = 1$, $\mu_0 = 0$, and $\tau_0^2 = 1$. We choose the sample mean $T(y) = \sum_1^n y_i/n$ as a discrepancy measure to check the fit of our model to $y_{obs}$ under both the prior and posterior predictive methods.

Under the prior predictive method we start by simulating a sample of prior parameters $\theta_j \sim N(\mu_0, \tau_0^2)$ for $j = 1,\ldots, 1000$, and then simulating prior predictive data sets $y_{rep}^{(j)} \sim N(\theta_j, \sigma^2)$, where the sample size of each is $n = 50$. We compute the measure $T$ on these to get a prior predictive sample $T(y_{rep}^{(j)})$, $j = 1,\ldots, 1000$, and also compute the observed value of the measure, $T(y_{obs})$. For a visual comparison, we plot these in the left side of figure 3.1. The prior predictive sample of $T$ is displayed in a histogram and the observed value is plotted as a vertical line. We can see very clearly that the value of $T(y_{obs}) \approx 19.2$ is in the extreme right-tail of the distribution. We calculate prior predictive $p$-value $p_{prior}$ to be $\sum_{j=1}^{1000} I(T(y_{rep}^{(j)}) \geq T(y_{obs}))/1000 = 0$. This essentially means there is an approximate zero probability of observing a value of $T$ at least as large as $T(y_{obs})$ within the prior predictive distribution. This strongly indicates that $y_{obs}$ was not generated under the considered model.

Under the posterior predictive method we start by computing the posterior mean and variance, $\tilde{\mu}_n$ and $\tilde{\tau}_n^2$ from (3.10). We then simulate a sample of posterior parameters $\tilde{\theta}_j \sim N(\tilde{\mu}_n, \tilde{\tau}_n^2)$ for $j = 1,\ldots, 1000$. Then we simulate posterior predictive data sets $\tilde{y}^{(j)} \sim N(\tilde{\theta}_j, \sigma^2)$ and lastly compute a posterior predictive sample of the measure, $T(\tilde{y}^{(j)})$, $j = 1,\ldots, 1000$.

We display these in a histogram in the right side of figure 3.1 along with the value of $T(y_{obs})$ plotted with a vertical line. The value of $T(y_{obs})$ is in the right-tail of the posterior predictive distribution. We compute the posterior predictive $p$-value $p_{post}$ as
Figure 3.1 Histograms of the prior predictive and posterior predictive samples of the discrepancy measure \( T(y) = \sum^n y_i / n \) simulated under the model \( y_i \sim N(\theta, 1) \); \( \theta \sim N(0, 1) \), along with the observed value \( T(y_{obs}) \) graphed as a vertical line. The left side shows the histogram of the prior predictive sample of \( T \) and the right side shows the histogram of the posterior predictive sample.

\[
\sum_{j=1}^{1000} I(T(\tilde{y}^{(j)}) \geq T(y_{obs}))/1000 = 0.029.
\]
So, the probability of observing a value of \( T \) at least as large as \( T(y_{obs}) \) in the posterior predictive distribution is approximately 2.9%. At a significance level of 5%, this is sufficient evidence to indicate that \( y_{obs} \) was not generated under the model.

### 3.2.4 Criticisms of the Prior Predictive and Posterior Predictive Methods

Many Bayesian statisticians have been hesitant to utilize prior predictive model diagnostics because the computed values of \( p_{prior} \) can depend critically on the model’s prior density (Johnson 2007). A very good model with a poor choice of prior distribution may be shown to not fit the data well. We note that posterior predictive diagnostics can be influenced by the prior distribution, but less so than with prior predictive.

A common concern about the posterior predictive \( p \)-value \( p_{post} \) is its double use of the
data. First, the data is used to compute the posterior distribution of the model parameters from which draws are used to replicate data from the posterior predictive distribution and second, to compute the value of $p_{post}$. \cite{BayarriBerger1999} give an example showing some problems because of this. In their example, they assume the null model to be $y_i \sim N(0, \sigma^2)$ and use discrepancy measure $T(y) = |\bar{y}|$. They showed that $p_{post}$ is always a positive constant even if $|\bar{y}| \to \infty$, and that this can be traced to the fact that the observed data is involved in the computation of both the posterior and the $p$-value.

Another concern about the posterior predictive $p$-value has to do with its distribution. In a frequentist setting, a key property that the $p$-value in (3.1) has is that under the null model its distribution is $U(0, 1)$, for all $\theta$ \cite{BayarriBerger2000}. This is a desirable property since it allows for a common interpretation across problems \cite{BayarriBerger1999}. Also, it is useful for determining a meaningful $\alpha$-level when doing model checking. In a frequentist setting, putting the significance level at $\alpha$ ($0 \leq \alpha \leq 1$) means that they can expect the $p$-value to be less than $\alpha$ about 100$\alpha\%$ of the time, i.e. $P[p \leq \alpha|M] \leq \alpha$. The prior predictive $p$-value $p_{prior}$ from (3.3) has a $U(0, 1)$ distribution under the prior predictive distribution $f(y)$ of (3.2) \cite{BayarriBerger2000}. The posterior predictive $p$-value however, has been shown not to be uniformly distributed, even asymptotically. \cite{Robinsetal2000} showed this and \cite{BayarriBerger2000} showed several examples illustrating this. \cite{Robinsetal2000} shows that distributions of posterior predictive $p$-values are concentrated more around 1/2 than are uniform. For a given model $M$ and $\alpha$-level then, this means that $P[p \leq \alpha|M] < \alpha$. That is, posterior predictive $p$-values may behave conservatively.

We show an example where the posterior predictive $p$-value is not uniformly distributed using the model $H_0 : y_i \sim N(\theta, \sigma^2) ; \theta \sim N(\mu_0, \tau_0^2)$ for $i = 1, \ldots, 50$, from section 3.2.3 and the measure $T(y) = \sum_1^n y_i/n$. We start by (1) generating an “observed” set of data $y_{obs}$ under the model $H_0$ and (2) compute the value of $p_{post}$ following the same steps taken to compute the posterior predictive $p$-value in the example in section 3.2.3. To see how the values of $p_{post}$ are distributed under $H_0$, we repeat this process a total of 1,000 times. To
Figure 3.2 Histograms of the prior predictive and posterior predictive \( p \)-values computed from the samples of the discrepancy measure \( T(y) = \sum_{i=1}^{n} y_i/n \) simulated under the model \( y_i \sim N(\theta, 1); \ \theta \sim N(0, 1) \), along with the observed value \( T(y_{obs}) \) graphed as a vertical line. The left side shows the histogram of the prior predictive sample of \( T \) and the right side shows the histogram of the posterior predictive sample.

see the distribution of the prior predictive \( p \)-value, for each of the 1,000 generated observed data sets we also computed the value of \( p_{prior} \) as done in section \[3.2.3\]. We display both of these distributions in figure \[3.2\].

The 1,000 values of \( p_{prior} \) are displayed in a histogram in the left side of figure \[3.2\]. We see that this is fairly uniform over the interval \([0,1]\). The values range from 0.0 to 1.0 with 51 values being less than 0.05. This means that the approximate type I error rate is 5.1% when the threshold level is 5.0%. On the other hand, we have the 1,000 values of \( p_{post} \) displayed in a histogram in the right side of figure \[3.2\]. Here we see the exact issues mentioned above by [Robins et al. (2000)]. The \( p \)-values are far from uniform, but instead are concentrated around 0.50. In fact, the values only range from about 0.36 to 0.64. So, the type I error rate is 0 at a threshold level of 5.0%. This shows that the values of \( p_{post} \), at least under this example, behave very conservatively.

Bayarri and Berger (1999) propose alternative methods that produce \( p \)-values that are asymptotically uniformly distributed under the null hypothesis. They propose calculating \( p \)-values from the *partial posterior-predictive* distribution which they term a *partial posterior-***
predictive p-value or $p_{ppost}$. They also propose a conditional predictive p-value or $p_{cpred}$. Bayarri and Berger (2000) illustrated the use of these and Bayarri and Castellanos (2007) showed examples comparing their performance to posterior-predictive p-values. The difficulty with these methods is that they are not easy to implement in cases where the model is even moderately complicated.

In the next two sections we introduce two alternative Bayesian model checking methods that offer some advantages over the previous methods. One is that they allow for discrepancy measures to be defined as functions of parameters only, and do not depend on the data. This can be beneficial in many aspects of Bayesian model checking. To test aspects of fit at higher levels of a hierarchical model, or to check for outlying parameter values, it may be beneficial to use discrepancy measures that can be defined at levels other than the data generating level. Also, as we will see in chapter 5, there may be aspects of model fit that can be checked by looking solely at the posterior values of parameters that are produced by fitting a given model. In this case, it is necessary to use discrepancy measures that are functions of these parameters alone. Finally, another advantage is that these two methods may be less conservative than the posterior predictive method.

3.3 Alternative Bayesian Model Checking Methods:

Prior Predictive Posterior Simulation (PPPS) Based Approach

In this section, we will introduce the Prior Predictive Posterior Simulation (PPPS) method of Dey et al. (1998). We will start by introducing the method, giving some of the motivations for it and the types of discrepancy measures that can be used. Then we will discuss the exact step by step details of how model checking is done with this procedure. To help illustrate the steps in the procedure, we finish the section with an example.
3.3.1 Introduction to the PPPS Approach

Dey et al. (1998) proposed an alternative model checking procedure to prior and posterior predictive checking. This new approach is entirely simulation based and can be used with any model in which we have a full model specification and can simulate draws from the posterior distribution. Dey et al. (1998) originally proposed this method for checking aspects of fit in hierarchical models, which prior or posterior predictive checking may not be suitable for.

In hierarchical models there are higher levels, other than the data generating level, that entail assumptions on how the parameters are distributed. One may be interested in checking aspects of fit at these higher levels, and in doing so may need to use discrepancy measures that are functions of the parameters only and not the data \( y \). In using the prior predictive or posterior predictive approach the discrepancy measures used depend on \( y \), and so could not be used for model checking at higher levels in the model.

We will now give an example for using this approach. Consider the hierarchical model

\[
\begin{align*}
\text{Level 1} & \quad y_i &\sim & N(\theta_i, \sigma^2), & i = 1, \ldots, n \\
\text{Level 2} & \quad \theta_i &\sim & N(\mu, \tau^2) \\
\text{Level 3} & \quad \mu &\sim & N(0, 1), & \sigma^2, \tau^2 \text{ known.}
\end{align*}
\]

We choose this model as a simple example to illustrate the ideas discussed in this section. From a practical standpoint, a model like this could be used to model data over different groups. In Level 1, \( y_i \) is an observation in group \( i \) with mean \( \theta_i \) and variance \( \sigma^2 \). In Level 2, the group means \( \theta_i \) are centered around an overall mean \( \mu \) with variance \( \tau^2 \), and Level 3 gives the prior distribution of the group mean \( \mu \).

For example, Draper (2011) discussed a hierarchical model very similar to (3.11) to model data from different experiments testing the effect of aspirin on heart attacks. In this, for \( n \) different randomized controlled experiments comparing the use of aspirin and placebo of patients following a heart attack, the observed difference in mortality rate between the
aspirin group and placebo group is recorded as $y_i$, $i = 1, \ldots, n$. The first level of the model treats the observed differences $y_i$ like random draws from normal distributions with mean $\theta_i$ and variance $\sigma^2$. The mean $\theta_i$ which is normally distributed with mean $\mu$ and variance $\tau^2$, can be viewed in this example as an underlying random effect due to relevant differences in patient groups and treatment protocols in each $i^{th}$ experiment.

We can check for outliers at the different levels with discrepancy measures say, $d_1(y, \theta) = y_i - \theta_i$ and $d_2(\theta, \mu) = \theta_i - \mu$ (Dey et al. 1998, p.331). To think of “outliers” at different levels of the model may seem strange. In this particular case, an outlier at Level 1 would be an observed value from a group that is extreme. Under the aspirin example above, this would represent an observed difference in mortality rates that is unexpectedly large or small under the model. An outlier at Level 2 would be a group mean that is extreme under the model. For example, this might represent a significant effect played by the differences in patient cohorts given the aspirin or placebo. In either case, trying to indicate these outliers is what motivates the residual based measures $d_1$ and $d_2$ for Level 1 and Level 2, respectively. We can also define discrepancy functions at Level 1 and 2 such as $d_1(y, \theta) = \sum_{i=1}^{n} (y_i - \theta_i)^2$ or $d_2(\theta, \mu) = \sum_{i=1}^{n} (\theta_i - \mu)^2$ to check to global fit of the model at that level.

We should reiterate here that the above discrepancy measures are used to check aspects of fit of model (3.11). Different models may require using different discrepancy measures. We are using the model in (3.11) as an example, but the following model checking procedure can be used for non-hierarchical models as well. All that is needed is the full model specification, and that for a given data set one be able to simulate draws from the posterior under the model. The full model specification (i.e. have a specified proper prior distribution) is required because, as we will see we will need to be able to simulate data from the prior predictive distribution.
3.3.2 Explanation of the PPPS Model Checking Procedure

We now explain the model checking procedure. Let $d$ be a discrepancy measure for a particular aspect of fit of the model. If we have the ability to compute the posterior distribution of the parameters given a set of observed data $\mathbf{y}_{\text{obs}} = (y_1^{(0)}, \ldots, y_n^{(0)})^t$, and draw a posterior sample of the parameters, we can use these to compute a posterior sample of the measure $d$. For example, using the measure $d_{1,i}(y_i, \theta_i) = y_i - \theta_i$, we would simulate, say $L$ draws from the marginal posterior distribution of $\theta_i$, given $\mathbf{y}_{\text{obs}}$. With this posterior sample $\{\tilde{\theta}_i^{(l)}\}_{l=1}^L$ we can compute a posterior sample of the measure, $d_{1,i}^{(l)} = y_i^{(0)} - \tilde{\theta}_i^{(l)}$, $l = 1, \ldots, L$.

What we have now is a sample from the posterior distribution of $d_{1,i}$ given the observed data. The question now is what can we reference this sample to to indicate to us if the values are extreme? In the posterior predictive method, we would use the posterior predictive distribution of $d_{1,i}$ by simulating a sample of posterior predictive data $\{\tilde{y}_i^{(l)}\}_{l=1}^L$ and computing a posterior predictive sample of the measure, $\{d_{1,i}(\tilde{y}_i^{(l)}, \tilde{\theta}_i^{(l)})\}_{l=1}^L$.

Under this new procedure, Dey et al. (1998) note that we can simulate a data set $\mathbf{y}_{\text{rep}} = (y_{\text{rep},1}, \ldots, y_{\text{rep},n})^t$ under the model and follow the same steps as above to simulate a posterior sample of the measure given this data. That is, we can simulate a data set $\mathbf{y}_{\text{rep}}$ from the prior predictive distribution under the proposed model and draw a sample $\{\theta_i^{*(l)}\}_{l=1}^L$ from the marginal posterior given $\mathbf{y}_{\text{rep}}$. Then, we can use this to compute a posterior sample $\{d_{1,i}(\mathbf{y}_{\text{rep},i}, \theta_i^{*(l)})\}_{l=1}^L$ of the measure. This gives us a sample from the posterior distribution of $d_{1,i}$ given the data $\mathbf{y}_{\text{rep}}$ arising under the model. This is what Dey et al. (1998) propose to compare the posterior sample $\{d_{1,i}(y_i^{(0)}, \tilde{\theta}_i^{(l)})\}_{l=1}^L$ arising from the observed data. However, here we have only simulated one prior predictive data set $\mathbf{y}_{\text{rep}}$, yielding only one posterior to use for comparison.

In order to have a good idea of how “extreme” the posterior values of $d_{1,i}$ given the observed data are, Dey et al. (1998) suggest repeating the above for many replicated data sets $\mathbf{y}_{\text{rep}}$ under the model. For say, $M$ replicated data sets, we would then use these to
obtain $M$ different posterior samples of the measure. We can then compare all these to the posterior sample calculated from the observed data.

To summarize, for a given measure $d$, the approach of Dey et al. (1998) is to compare the posterior distribution of $d$ given $y_{obs}$ to posterior distributions of $d$ that arise given data $y_{rep}$ generated under the model. The idea is that if the observed data was generated under the model, then the corresponding posterior distribution of $d$ should be similar to those posteriors of $d$ given $y_{rep}$. The way of comparing these distributions is by simulating representative samples, as discussed above, and comparing these. We note here that we call this procedure the prior predictive posterior simulation (PPPS) method, since we are simulating from posterior distributions that arise from data generated under the prior predictive distribution (Yuan and Johnson 2011).

The next logical question is, how do we compare the posterior sample of $d$ given $y_{obs}$ with the posterior samples of $d$ given the replicated data sets? We could use summary statistics to summarize the respective samples, and compare these. For example, we could compute the sample mean of the sample of $d$ given $y_{obs}$, and compare this to the sample means of $d$ given the replicated data sets. However, some information about a given sample is lost when computing a single summary measure. Dey et al. (1998) recommend computing the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles of each sample and using these for comparison. Next, we will outline the procedure that is used in Dey et al. (1998).

The comparison procedure for the posterior samples as outlined in Dey et al. (1998) is given here.

Set Up:
Let $y^{(0)} = y_{obs}$ and simulate $M$ data sets $y_{rep}^{(m)}$, $m = 1, \ldots, M$ from the prior predictive distribution. From the $M$ replicated data sets $y_{rep}^{(m)}$ we have a corresponding set of $M$ posterior distributions, $\pi(d|y_{rep}^{(m)})$. For each $m$, we simulate a sample from the posterior $\pi(d|y_{rep}^{(m)})$ by drawing a posterior sample of size $L$ of the necessary parameters and calculating $L$ values of the measure $d$. Denote these as $d_{l}^{(m)}$, $l = 1, \ldots, L$. We also simulate a sample
of size $L$ from the posterior of $d$ given the observed data, $\pi(d|y^{(0)})$. Denote these as $d^{(0)}_l$, $l = 1, \ldots, L$. If the discrepancy $d$ is continuous then we compare the sample $\{d^{(0)}_l\}_{l=1}^L$ with the samples $\{d^{(1)}_l\}, \ldots, \{d^{(M)}_l\}$ by:

1. For the sample $\{d^{(0)}_l\}_{l=1}^L$, compute the vector of five quantiles $q^{(0)} = (q_{0.05}^{(0)}, q_{0.25}^{(0)}, q_{0.50}^{(0)}, q_{0.75}^{(0)}, q_{0.95}^{(0)})^t$.

2. For each sample $\{d^{(m)}_l\}_{l=1}^L$, $m = 1, \ldots, M$, compute the vector of five quantiles: $q^{(m)} = (q_{0.05}^{(m)}, q_{0.25}^{(m)}, q_{0.50}^{(m)}, q_{0.75}^{(m)}, q_{0.95}^{(m)})^t$.

3. Compute the mean vector over the $q^{(m)}$, $m = 1, \ldots, M$, from (2), giving $\bar{q} = (\bar{q}_{0.05}, \bar{q}_{0.25}, \bar{q}_{0.50}, \bar{q}_{0.75}, \bar{q}_{0.95})^t$.

4. For $q^{(0)}$ and each of $q^{(m)}$, compute the Euclidean distance

$$e^{(0)} = \|q^{(0)} - \bar{q}\| = \left(\sum_i (q^{(0)}_i - \bar{q}_i)^2\right)^{1/2}, \quad (3.12)$$

and

$$e^{(m)} = \|q^{(m)} - \bar{q}\| = \left(\sum_i (q^{(m)}_i - \bar{q}_i)^2\right)^{1/2}, \quad (3.13)$$

for $i = 0.05, 0.25, 0.50, 0.75, 0.95$.

5. With the resulting set of $M + 1$ Euclidean distances, a one-sided upper tail 0.05 level Monte Carlo test is performed. Specifically, we consider the null hypothesis $H_0$: the level in question of the posited model is true. We compute the empirical $p$-value by

$$\hat{P}(e^{(m)} \geq e^{(0)} | H_0) = \frac{\sum_{i=1}^M I(e^{(m)} \geq e^{(0)})}{M + 1}. \quad (3.14)$$

If this is less than the threshold level of 0.05, then this is taken as evidence that the model specification for that level may not be a good fit with the observed data.

As mentioned, Dey et al. (1998) is using the quantiles $q^{(0)}$ and $q^{(m)}$, $m = 1, \ldots, M$, as
summary measures for the respective posterior samples \( \{d_i^{(0)}\}_{i=1}^L \) and \( \{d_i^{(1)}\}, \ldots, \{d_i^{(M)}\}\}_{i=1}^L \). As a way to compare the quantile vector \( q^{(0)} \) to \( q^{(m)} \), \( m = 1, \ldots, M \), we compute the Euclidean distance of each from the mean quantile vector \( \bar{q} \). We then compare the Euclidean distance \( e^{(0)} \) to the Euclidean distances \( e^{(m)} \), \( m = 1, \ldots, M \), by computing the \( p \)-value in (3.14), where small values indicate lack of fit. The rationale behind this is if the proportion of distances \( e^{(m)} \) greater than \( e^{(0)} \) is small, then this indicates that \( e^{(0)} \) is large relative to the other distances. This implies that the quantiles of the sample \( \{d_i^{(0)}\}_{i=1}^L \) do not agree with the quantiles of the samples \( \{\{d_i^{(1)}\}, \ldots, \{d_i^{(M)}\}\}_{i=1}^L \), that were computed from data replicated under the model.

If the discrepancy measure \( d \) is discrete, then the calculation of the quantiles in steps 1 and 2 above is replaced by creating five bins for the values of \( d \) and calculating the bin proportions. Step 3 is analogously replaced with finding the average bin proportions over the samples drawn from the posteriors based on the replicated data sets. Steps 4 and 5 may be performed the same, taking the Euclidean distances between the vectors of bin proportions and the vector of averages along with running a one-sided Monte Carlo test on the resulting set of \( M + 1 \) distances.

### 3.3.3 Example Illustrating the PPPS Model Checking Method

We will give an example using the PPPS method to check the fit of values at the first level of the model in (3.11). The discrepancy measures we will use are \( d_{1,i}(y_i, \theta_i) = y_i - \theta_i \), \( i = 1, \ldots, n \), where we will take \( n = 10 \). To see that the model checking procedure is useful, we will generate the observed data \( y^0 = (y_1^{(0)}, \ldots, y_{10}^{(0)})^t \) so that the first value \( y_1^{(0)} \) is extreme under the model and the other observations \( y_2^{(0)}, \ldots, y_{10}^{(0)} \) according to the model. Our goal here is to see if the PPPS procedure can indicate the data value \( y_1^{(0)} \) as being aberrant under the model.

The data is generated by simulating \( \mu \sim N(0, 1) \) and then simulate \( \theta_i \sim N(\mu, \tau^2) \), \( i = 1, \ldots, 10 \) where \( \tau^2 = 1 \). We then simulate \( y_1^{(0)} \sim N(\theta_1, 2\sigma^2) \) and \( y_i^{(0)} \sim N(\theta_i, \sigma^2) \),
We draw samples \( \{ \tilde{\theta}_i^{(l)} \}_{l=1}^{1000} \), for \( i = 1, \ldots, 10 \), from the marginal posterior given \( y^0 \), and use these to compute posterior samples of the measure, \( \{ d_{1,i}^{(0)}(y_i^{(0)}, \tilde{\theta}_i^{(l)}) \}_{l=1}^L \). The vector of the five quantiles \( q_i^{(0)} \), \( i = 1, \ldots, 10 \), is then computed for these samples.

For each \( i = 1, \ldots, 10 \), we iterate this next part for \( M = 500 \) times. For \( m = 1, \ldots, 500 \), we simulate data \( y_{\text{rep}}^{(m)} = (y_1^{(m)}, \ldots, y_{10}^{(m)})^t \) from the prior predictive distribution. We then simulate draws \( \{ \tilde{\theta}_i^{(m)(l)} \}_{l=1}^L \) from the marginal posterior distribution given \( y_{\text{rep}}^{(m)} \), and use these to compute a posterior sample of the measure, \( \{ d_{1,i}^{(m)}(y_i^{(m)}, \tilde{\theta}_i^{(m)(l)}) \}_{l=1}^L \). We then compute the vector of five quantiles \( q_i^{(m)} \). After iterating this process 500 times we end up with a set of 500 quantile vectors \( q_i^{(m)} \), \( m = 1, \ldots, 500 \). Lastly, for each \( i = 1, \ldots, 10 \), we compute the mean vector \( \bar{q}_i \) over the 500 quantile vectors.

Before continuing, it might be useful to recap what we have done so far. Again, the approach of the PPPS method is to compare posterior samples of a discrepancy measure \( d \) that are computed from data sets simulated under the model (prior predictive distribution) to a posterior sample of \( d \) computed given the observed data \( y_{\text{obs}} \). If \( y_{\text{obs}} \) was generated under the model, then we would expect the corresponding posterior sample of \( d \) to be similar to the others. If the posterior sample of \( d \) given \( y_{\text{obs}} \) is not similar, this casts doubt on \( y_{\text{obs}} \) having been generated under the model. The way we are choosing to compare the posterior sample of \( d \) given \( y_{\text{obs}} \) to the others is by using the Euclidean distances of the quantile vectors as described above.

We try to illustrate this in figure 3.3. In figure 3.3 we plot the posterior sample densities of \( d_{1,i}^{(0)} \) and \( d_{1,i}^{(m)} \) for \( i = 1, 4, 6 \) and 8. Out of the \( M = 500 \) posterior samples \( d_{1,i}^{(m)} \), we randomly selected 50 to plot sample densities of (light color) to compare to the sample density of \( d_{1,i}^{(0)} \) (dark color). In each plot, the 5 values in the mean quantile vector \( \bar{q}_i \) are represented by the 5 vertical lines, respectively. The 5 quantiles of \( q_i^{(0)} \) are plotted as bold points on the horizontal axis for comparison.

First off, we see that the posterior sample density of \( d_{1,1}^{(0)} \) is shifted to the right of the
Figure 3.3 Posterior sample densities plotted for $d_{1,i}^{(0)}$ and $d_{1,i}^{(m)}$ for $i = 1, 4, 6$ and 8. Out of the $M = 500$ posterior samples $d_{1,i}^{(m)}$, we randomly selected 50 to plot sample densities of (light color) to compare to the sample density of $d_{1,i}^{(0)}$ (dark color). In each plot, the 5 values in the mean quantile vector $\bar{q}_i$ are represented by the 5 vertical lines, respectively. The 5 quantiles of $q_i^{(0)}$ are plotted as bold points on the horizontal axis for comparison.
other posterior sample densities of $d_{1,1}^{(m)}$ (top left). The other sample densities of $d_{1,i}^{(0)}$, $i = 4, 6,$ and 8, appear to be very typical in reference to the sample densities of $d_{1,1}^{(m)}$. Following from this, we see that the 5 quantiles of $q_1^{(0)}$ are respectively larger than the mean quantile values of $q_1$. In the other three plots, the values of $q_i^{(0)}$ are very close, respectively, to the mean values of $q_i$. This is what we would expect to see, since the data value $y_1^{(0)}$ was generated to be outlying under the model, and the values $y_{4}^{(0)}$, $y_6^{(0)}$, and $y_8^{(0)}$ were not.

As mentioned, we quantify the “difference” between the posterior samples of $d_{1,i}^{(0)}$ and $d_{1,i}^{(m)}$, $i = 1, \ldots, 10$, by comparing the Euclidean distances $e_i^{(m)}$ computed between each of $q_i^{(m)}$ and $q_i$, and the Euclidean distance $e_i^{(0)}$ between $q_i^{(0)}$ and $q_i$. A $p$-value is then calculated by looking at the proportion of Euclidean distances $e_i^{(m)}$ that are greater than $e_i^{(0)}$.

We show the calculated $p$-values in table 3.1. Once again, calculated $p$-values less than 0.05 indicate that the posterior sample of $d_{1,i}$ computed based on the observed data was not in agreement with those computed using data generated under the model. We see that the $p$-value corresponding to the measure $d_{1,1}$ computed for the observation $y_1^{(0)}$ is less than 0.05. This indicates that the observed value of $y_1^{(0)}$ is outlying under the model, which indeed it was. The rest of the $p$-values are all greater than 0.05, which means that the measure indicates no discrepancy between the values $y_2^{(0)}, \ldots, y_{10}^{(0)}$ and the model. This is what we would expect since these values were generated under the model.

| Empirical $p$-values for $d_{1,i}(y_i, \theta_i) = y_i - \theta_i$ |
|--------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| $d_{1,1}$ ($y_{11}^{(m)}$) $e_i^{(m)}$ | $d_{1,1}$ | $d_{1,2}$ | $d_{1,3}$ | $d_{1,4}$ | $d_{1,5}$ | $d_{1,6}$ | $d_{1,7}$ | $d_{1,8}$ | $d_{1,9}$ | $d_{1,10}$ |
| $P(e_i^{(m)} \geq e_i^{(0)})$ | 0.010 | 0.872 | 0.906 | 0.848 | 0.441 | 0.733 | 0.581 | 0.545 | 0.555 | 0.749 |

Table 3.1 Empirical $p$-values computed using the discrepancy measure $d_{1,i}(y_i, \theta_i) = y_i - \theta_i$, $i = 1, \ldots, 10$, to check for outlying values at Level 1 of model (3.11). The observed data was simulated as $y_1^{(0)} \sim N(\theta_1, 2\sigma^2)$ which should produce a more extreme value than the regular model, and $y_i^{(0)} \sim N(\theta_i, \sigma^2)$, $i = 2, \ldots, 10$. Calculated $p$-values less than 0.05 indicate that the posterior sample of $d_{1,i}$ computed based on the observed data was not in agreement with those computed using data generated under the model.

This method offers promise of being a more powerful method for doing model diagnostics.
Unlike the posterior predictive approach which can be conservative and relies on a “double use of the data”, with the PPPS method we are directly comparing the posterior arising from the observed data with posteriors expected from data under the model. Of course, this method is computationally expensive. For example, if we are using MCMC to draw samples from the posterior distribution then in the posterior predictive approach we would use one run of our MCMC algorithm to get a posterior sample and use this to compute the posterior predictive distribution. In the PPPS method, along with one run of the MCMC algorithm to get a sample from the posterior under the observed data, we need to run this $M$ more times for each $y_{rep}$ as mentioned above. So, the overall process will take $M$ times longer.

### 3.4 Alternative Bayesian Model Checking Methods: 
Assessment Using Pivotal Discrepancy Measures (PDMs)

In this section we will describe a method of Bayesian model checking that involves using pivotal discrepancy measures (PDMs). The idea of this method began in Johnson (2004) and was formalized in Johnson (2007) and Yuan and Johnson (2011). We will start by introducing what a PDM is and then, for the reader’s education, discussing the background of the method. The method using PDMs relies on two important theoretical results, and these will be discussed and illustrated with an example. Lastly, we will discuss the actual method by which Johnson (2007) and Yuan and Johnson (2011) propose to use PDMs for model checking.

#### 3.4.1 Introduction to the PDM Method

Johnson (2007) introduced a method of Bayesian model checking based on using discrepancy measures that are pivotal quantities. A pivotal quantity is defined in Johnson (2007) as a function $T(y, \theta)$ that has an invariant distribution when evaluated at the true (data generating) parameter $\theta_0$. This can be interpreted to be a pivotal quantity in the usual sense...
that the distribution of $T(y, \theta_0)$ does not depend on the parameters. \textcite{Johnson2007} and \textcite{YuanJohnson2011} show that if one can define a discrepancy measure that is pivotal, then one will also be able to find a known reference distribution for this measure.

An advantage that we will see with this method is, like under the PPPS method, one is able to define measures that are functions of the data and parameters, or just parameters themselves. The results that we will implement are mainly found in \textcite{Johnson2007} and \textcite{YuanJohnson2011}. As mentioned, the basis for these results is given in \textcite{Johnson2004} and so we will discuss these first here.

### 3.4.2 Background of PDMs

\textcite{Johnson2004} first proposed a modification of the classic $\chi^2$ statistic used in goodness of fit tests. This produces a Bayesian analogue to the $\chi^2$ goodness-of-fit test. This was shown for both continuous and discrete data, but we will show an example here with the continuous case. Let $y = (y_1, \ldots, y_n)^t$ be a vector of real-valued, continuous, identically distributed, conditionally independent observations drawn from sampling density $f(y|\theta)$ where $\theta$ is an $s$-dimensional parameter vector, $s \geq 1$, with prior distribution $p(\theta)$. Let $\tilde{\theta}$ denote a sample from the posterior distribution based on $y$.

The interval $[0, 1]$ is then partitioned into $K$ bins with $0 = a_0 < a_1 < \ldots < a_{K-1} < a_K = 1$ and let $p_k = a_k - a_{k-1}$, $k = 1,\ldots,K$. From here, we look at the values of $F(y_i|\tilde{\theta})$, $i = 1,\ldots,n$, where $F(\cdot|\tilde{\theta})$ is the cumulative distribution function of $Y$, and count how many fall into each each of the $K$ bins $(a_{k-1},a_k]$. Denote these counts by $m_k(\tilde{\theta})$ where

$$m_k(\tilde{\theta}) = \sum_{i=1}^{n} I \left( F(y_i|\tilde{\theta}) \in (a_{k-1},a_k] \right)$$

where $I(\cdot)$ is an indicator function. What \textcite{Johnson2004} termed as a Bayesian $\chi^2$ statistic is then defined as

$$R(\tilde{\theta}) = \sum_{k=1}^{K} \left[ \frac{(m_k(\tilde{\theta}) - np_k)}{\sqrt{np_k}} \right]^2. \quad (3.15)$$
It is shown that under certain regularity conditions that the statistic \( R(\tilde{\theta}) \) converges to a \( \chi^2 \) distribution with \( K - 1 \) degrees of freedom as the sample size of the observed data \( n \) goes to infinity. To avoid future confusion, we make the fine points of the claim clear. Provided we have observed data \( y \) that was generated under the model with sample size sufficiently large, then \( R(\tilde{\theta}) \) asymptotically follows a \( \chi^2 \) distribution when one value of \( \tilde{\theta} \) is drawn from the posterior \( \pi(\tilde{\theta} | y) \) induced by the observed data. The result does not necessarily hold for multiple values of \( \tilde{\theta} \) drawn from the posterior \( \pi(\tilde{\theta} | y) \) based on the same observed data. The importance of this result is that it motivates the work of [Johnson (2007)] which proves a theorem that shows a similar result holds for any measure that is a *pivotal quantity*.

### 3.4.3 Some Important Theoretical Results for PDMs

As mentioned, [Johnson (2007)] defined a *pivotal quantity* as a function \( T(y, \theta) \) that has an invariant distribution when evaluated at the true (data generating) parameter \( \theta_0 \). This can be interpreted to be a pivotal quantity in the usual sense that the distribution of \( T(y, \theta_0) \) does not depend on the parameter(s) \( \theta \). For example, look at the first level of the hierarchical model given in (3.11). Since \( y_i \sim N(\theta_i, \sigma^2) \), then defining \( d_1(y, \theta) = (y_i - \theta_i)/\sigma \) has a \( N(0, 1) \) distribution which means \( d_1(y, \theta) \) is a pivotal quantity. In relation to these pivotal quantities, [Johnson (2007)] proved the following theorem,

**Theorem 1.** Let \( T(Y, \theta) \) denote a pivotal quantity, and suppose that \( \theta_0 \) is a random vector drawn from density \( \pi \). Given \( \theta_0 \), let \( Y \) denote a random vector drawn from density \( f(y|\theta_0) \), and let \( \theta_Y \) denote a parameter vector drawn from the posterior distribution on \( \theta \) given \( Y \). Then \( T(Y, \theta_Y) \) and \( T(Y, \theta_0) \) are identically distributed.

To help the explanation, let \( d(Y, \theta) \) denote a pivotal quantity, let \( \theta_0 \) be the true parameter value for which data \( y \) is generated, and let again \( \tilde{\theta} \) be a single value of the parameter drawn from the posterior based on \( y \). What Theorem 1 tells us then is that the distributions of \( d(y, \theta_0) \), and \( d(y, \tilde{\theta}) \) are the same. The application of this result to model checking is very
easy. Even though the true parameter value $\theta_0$ is unknown we can take a draw $\tilde{\theta}$ from the posterior to evaluate the measure, and use the known distribution of $d(y, \theta_0)$ as a reference distribution. Then, we can use the values of $d$ evaluated at posterior draws of the parameter to check against the reference distribution. If there is discordancy, then this implies model lack of fit.

Theorem 1 deals with discrepancy measures $T(y, \theta)$ that are functions of the data and parameters. Yuan and Johnson (2011) extends this to consider pivotal quantities that are functions of just parameters. To include discrepancy measures that are either functions of data and parameters, or just parameters, the term pivotal discrepancy measure (PDM) was introduced. We will use this same term throughout this paper. The following lemma was proved,

**Lemma 1.** Suppose that $T(y, \theta^0)$ is a PDM distributed according to $F$. If $\tilde{\theta}$ is drawn from the posterior on $\theta$ given $y$, then $T(y, \tilde{\theta})$ is also distributed according to $F$.

The way Lemma 1 is worded makes it sound like it is dealing with PDMs that are functions of the data and parameters, and not parameters alone. It is unclear why it is worded this way, but the proof given in Yuan and Johnson (2011) is for the case of PDMs relying on parameters only.

Again, as an example, look at the second level of the hierarchical model in (3.11). We have $\theta_i \sim N(\mu, \tau^2)$, and let us define $d_{2i}(\theta_i, \mu) = (\theta_i - \mu)/\tau$. This is a pivotal quantity having a $N(0, 1)$ distribution. Then according to Lemma 1 if we take a posterior draw $(\tilde{\theta}_i, \tilde{\mu})$, then the measure $d_{2i}(\tilde{\theta}_i, \tilde{\mu}) = (\tilde{\theta}_i - \tilde{\mu})/\tau$ has a $N(0, 1)$ distribution. We should note that for the rest of this chapter we will use the term reference distribution of a PDM to mean the distribution the measure follows when evaluated using the true value of the parameter(s).

To see an example that illustrates the results of Theorem 1 and Lemma 1 we perform the following simulation using the hierarchical model from (3.11), and the above defined PDMs, $d_{1i}(y_i, \theta_i)$ and $d_{2i}(\theta_i, \mu)$. For 1,000 iterations, we simulate data $y_i$, $i = 1, \ldots, 10$ from the model by (1) simulating $\mu_0 \sim N(0, 1)$, (2) simulating $\theta_{0i} \sim N(\mu_0, \tau^2)$ with $\tau^2 = 1$ and
(3) simulate \( y_i \sim N(\theta_{0i}, \sigma^2) \) with \( \sigma^2 = 4 \). We then compute the posterior and simulate one draw of \( \tilde{\theta} \) and \( \tilde{\mu} \) from their marginal posteriors \( \pi(\theta|y) \) and \( \pi(\mu|y) \). We calculate \( d_{1i}(y_i, \tilde{\theta}_i) \), and \( d_{2i}(\tilde{\theta}_i, \tilde{\mu}) \) for \( i = 1, \ldots, 10 \). So, for each iteration we obtain one value of \( d_{1i}(y_i, \tilde{\theta}_i) \) and \( d_{2i}(\tilde{\theta}_i, \tilde{\mu}) \), which according to Theorem 1 and Lemma 1, respectively, have a \( N(0,1) \) distribution. Iterating 1,000 times then should give us a sample of 1,000 \( N(0,1) \) random variables.

To check this, we plot the 1,000 simulated values of \( d_{1i}(y_i, \tilde{\theta}_i) \), and \( d_{2i}(\tilde{\theta}_i, \tilde{\mu}) \), \( i = 1, \ldots, 10 \), in normal quantile plots in figures 3.4 and 3.6, respectively. The top of each plot is labeled with its respective index for \( i = 1, \ldots, 10 \). The theoretical quantiles of the \( N(0,1) \) distribution are plotted on the \( x \)-axis and the sample quantiles of the simulated samples are plotted on the \( y \)-axis. The purpose of these plots is to see how close the simulated samples are to being from a \( N(0,1) \) distribution by seeing how close the sample quantiles are to the theoretical quantiles. The line \( y = x \) is graphed in each plot for reference. The closer the plotted points lie to this line, the closer the respective quantiles are to being equal.

Looking at figures 3.4 and 3.6 we see that the quantiles of the two simulated samples very closely match that of the \( N(0,1) \) distribution, in that the points all fall very close to the line \( y = x \) in each plot. We do see some deviation from the line \( y = x \) for very small or large quantiles (i.e. less than \(-2\), or greater than \(2\)), but the sample quantiles seem very close to the theoretical quantiles throughout the middle 95% of the \( N(0,1) \) distribution.

In figures 3.5 and 3.7 we plot Tukey mean-difference plots for the quantiles plotted in figures 3.4 and 3.6, respectively. The purpose of these plots is to give us an idea of how far the sample quantiles are from the theoretical quantiles. In each plot, the theoretical quantiles of the standard normal distribution are plotted on the \( x \)-axis and the difference between the sample and theoretical quantiles is plotted on the \( y \)-axis. That is, letting \( d_{1i}(r) \) be the \( r^{th} \) quantile of \( d_{1i}(y_i, \tilde{\theta}_i) \), and say, \( z(r) \) be the \( r^{th} \) standard normal quantile, we plot \( z(r) \) on the horizontal axis and \( d_{1i}(r) - z(r) \) on the vertical axis of figure 3.5. Likewise, letting
Figure 3.4 Normal quantile plots of the 1,000 simulated values of $d_{1i}(y_i, \tilde{\theta}_i) = (y_i - \tilde{\theta}_i)/\sigma$ for each $i = 1, \ldots, 10$. The theoretical quantiles of the $N(0, 1)$ distribution are plotted on the x-axis and the sample quantiles are plotted in the y-axis. The line $y = x$ is plotted for reference.

d_{2i(r)}$ be the $r^{th}$ quantile of $d_{2i}(\tilde{\theta}_i, \tilde{\mu})$ we plot $z(r)$ on the horizontal axis and $d_{2i(r)} - z(r)$ on the vertical axis of figure 3.7. The top of each of these plots is also labeled with its respective index for $i = 1, \ldots, 10$. A reference line at $y = 0$ is plotted to judge the differences plotted on the horizontal.

In both figures 3.5 and 3.7 the sample quantiles are close to the theoretical quantiles. We see that most of the differences plotted on the horizontal are between about $-0.15$ and $0.15$ for the quantiles between $-2$ and $2$. This shows that the sample quantiles are in fact very close to the theoretical quantiles, for at least the middle 95% of each sample.

This result now allows us to use measures that can depend on data, or parameters alone, as long as the measure is pivotal with a known reference distribution. We can use posterior draws of the parameters and check the corresponding value of $d$ against the reference distribution. Discordancy here will give evidence of outliers or model lack of fit. This method using PDMs has a large computational advantage over the PPPS approach and again can allow us to define discrepancy measures that do not rely on data, which the prior predictive
Figure 3.5 Tukey mean-difference plot for the quantiles plotted in figure 3.4. The theoretical quantiles $z(r)$, $r = 1, \ldots, 1000$, of the $N(0, 1)$ distribution are plotted on the $x$-axis and the difference $d_{1i}(r) - z(r)$ between the sample and theoretical quantiles is plotted on the $y$-axis. A reference line at $y = 0$ is plotted to judge the differences plotted on the horizontal.

Figure 3.6 Normal quantile plots of the 1,000 simulated values of $d_{2i}(\tilde{\theta}_i, \tilde{\mu}) = (\tilde{\theta}_i - \tilde{\mu})/\tau$ for each $i = 1, \ldots, 10$. The theoretical quantiles of the $N(0,1)$ distribution are plotted on the $x$-axis and the sample quantiles are plotted in the $y$-axis. The line $y = x$ is plotted for reference.
Figure 3.7 Tukey mean-difference plot for the quantiles plotted in figure 3.6. The theoretical quantiles $z(r)$, $r = 1, \ldots, 1000$, of the $N(0, 1)$ distribution are plotted on the $x$-axis and the difference $d_{2i(r)} - z(r)$ between the sample and theoretical quantiles is plotted on the $y$-axis. A reference line at $y = 0$ is plotted to judge the differences plotted on the horizontal.

and posterior predictive assessment methods both can not do.

The next question to ask is, for a given set of observed data $y_{obs}$ and a simulated posterior sample of a given PDM, how do we use this to perform model checking? That is, how can the posterior sample of the PDM be compared to the reference distribution to make inferences about the fit of the model to $y_{obs}$? Next we will discuss the methods proposed in [Johnson (2007)] and [Yuan and Johnson (2011)] for using PDMs for model checking.

### 3.4.4 Using PDMs for Model Checking

#### 3.4.4.1 Introduction

In our example in the previous section 3.4.3, for each of the 1,000 iterations a new data set $y_i$, for $i = 1, \ldots, 10$, was drawn and posterior values of $\tilde{\theta}_i$ and $\tilde{\mu}$ were simulated from their respective marginal posteriors given $y_i$. Under Theorem 1 and Lemma 1 the
corresponding values of \( d_1(y_i, \tilde{\theta}_i) \) and \( d_2(\tilde{\theta}_i, \tilde{\mu}) \) make up samples of 1,000 independent draws from a \( N(0, 1) \) distribution. In practice though we only have one set of observed data \( y_{\text{obs}} \) to work with. Under Theorem 1 and Lemma 1, we can draw values of \( \tilde{\theta} \) and \( \tilde{\mu} \) from their marginal posteriors \( \pi(\theta | y_{\text{obs}}) \) and \( \pi(\mu | y_{\text{obs}}) \) respectively, and each value of \( d_1(y_i(0), \tilde{\theta}_i) \) and \( d_2(\tilde{\theta}_i, \tilde{\mu}) \) can be viewed as a draw from a \( N(0, 1) \) distribution (provided \( y_{\text{obs}} \) was generated under the model). At this point, a hypothesis test in the classical sense could be performed by comparing the values of \( d_1(y_i(0), \tilde{\theta}_i) \) and \( d_2(\tilde{\theta}_i, \tilde{\mu}) \) to a specified critical value in the \( N(0, 1) \) distribution to judge their extremeness. However, as Johnson (2004) point out we do not gain much information from the posterior distribution by simulating only one value.

So, we could simulate a sample of say, 1,000 values of \( \tilde{\theta} \) and \( \tilde{\mu} \) from their marginal posteriors \( \pi(\theta | y_{\text{obs}}) \) and \( \pi(\mu | y_{\text{obs}}) \), respectively. Again, under Theorem 1 and Lemma 1 then, each computed value of \( d_1(y_i(0), \tilde{\theta}_i(l)) \) and \( d_2(\tilde{\theta}_i(l), \tilde{\mu}(l)), l = 1, \ldots, 1000 \), could be viewed as a draw from the \( N(0, 1) \) reference distribution (provided \( y_{\text{obs}} \) was generated under the model). From here, maybe we could perform a test for equal distributions, such as a Komorogorov-Smirnov or Cramer-von Mises test, between our sample and the reference distribution to check the adequacy of the model. However, these tests assume independent identically distributed samples.

As indicated in Johnson (2004), Johnson (2007), and Yuan and Johnson (2011), when multiple parameter values are drawn from the same posterior distribution (i.e. \( \pi(\theta | y_{\text{obs}}) \) or \( \pi(\mu | y_{\text{obs}}) \)) the resulting PDM values are not independent. So, each computed value of \( d_1(y_i(0), \tilde{\theta}_i(l)) \) and \( d_2(\tilde{\theta}_i(l), \tilde{\mu}(l)), l = 1, \ldots, 1000 \), could be viewed as a draw from the \( N(0, 1) \) distribution, but collectively these would form a dependent sample.

Well then, for a given set of observed data \( y_{\text{obs}} \) and a simulated posterior sample of a given PDM, how do we use this to perform model checking? Johnson (2007) proposed three methods for using PDMs for model checking. The first two are rather informal. The third method is the more formal, which we will implement when using this method on later chapters. We will briefly discuss the first two methods before discussing the third in more detail.
In all methods suppose we have a set of observed data $y_{obs}$ and a computed sample of posterior PDM values using simulated parameter values from the marginal posterior distribution given $y_{obs}$.

### 3.4.4.2 Informal Model Checking Methods using PDMs

The first method suggested is to simply do a graphical comparison between the posterior sample of the PDM and the reference distribution. For example, define the PDM $d_1(y, \theta) = \sum_{i=1}^{n} d_1(y_i, \theta_i)^2$ which has a $\chi^2_{10}$ reference distribution. Figure 3.8 shows the results of two cases where we computed posterior samples of 1,000 values of $d_1$ from data generated under the model (top), and data generated under a false model $N(\theta_i, 2.5^2\sigma^2)$ (bottom). A $\chi^2_{10}$ density curve is plotted in the top and bottom of figure 3.8 along with the histograms of the posterior samples. We can easily see that the bottom histogram does not match well with the reference distribution, indicating a lack-of-fit with the data. This method is good for a quick visual check, but is lacking if we want to make a more exact inference.

The second method is to specify some critical value in the reference distribution, and look at the proportion of PDM values that would result in a rejection of the null hypothesis of model fit. For example, in the top and bottom of figure 3.8 we overlay a horizontal line at the 95th percentile of the $\chi^2_{10}$ density curve. Since ideally each value of $d_1$ is a $\chi^2_{10}$ random variable, any value beyond the 95th percentile would be considered extreme. In a usual (classical) hypothesis test, a test statistic value this extreme would lead to a rejection of the null hypothesis, at the 5% level. We simply look at the whole posterior sample of $d_1$ and determine the percentage of values that are extreme. This may lead to a statement such as, 40% of the PDM values generated from the posterior distribution exceeded the 95th quantile of the reference distribution.

There is a fine point to understand here, that is we are not comparing the posterior distribution with the reference distribution. Since the single value of the PDM calculated
Figure 3.8 (Top) Histogram of a posterior sample of discrepancy values $d_1(y, \theta) = \sum_{i=1}^{10} (y_i - \theta_i)^2 / \sigma^2$ with data generated under the true model, i.e. $y_i \sim N(\theta_i, \sigma^2)$. (Bottom) Histogram of a posterior sample of $d_1$ values with $y_i \sim N(\theta_i, 2.5^2 \sigma^2)$. The true $\chi^2_{10}$ reference distribution density curve is overlaid in each for comparison.
from a single draw from the posterior follows the reference distribution, but also comes from
the posterior of the PDM, we are looking at the proportion of posterior values that we ‘could
have drawn’ that would have resulted in rejection.

3.4.4.3 Formal Model Checking Method using PDMs

Now we discuss the more formal method. Here we consider a PDM \( d(y, \theta) \) which follows
reference distribution \( F \), which is known. Since a posterior sample of \( d \) can be thought of
as a dependent sample from distribution \( F \), \cite{Johnson2007} and \cite{YuanJohnson2011}
proposed using bounds of order statistics of dependent, identically distributed random vari-
able. These bounds were derived in \cite{CarauxGascuel1992}. Let \( X_1, \ldots, X_n \) be a set of
\( n \) random variables that are identically distributed (and possibly dependent) with distribu-
tion function \( F \), and letting \( F_{X(r)} \) be the distribution function of the \( r^{th} \) order statistic from
the sample, then \cite{CarauxGascuel1992} showed that

\[
F_{X(r)}(x) \geq \max \left\{ 0, 1 - \frac{n(1 - F(x))}{n - r + 1} \right\} 
\]

(3.16)

\[
= \max \left\{ 0, \frac{nF(x) - r + 1}{n - r + 1} \right\}.
\]

To apply this to a posterior sample of the PDM \( d \), let \( D(r) \) denote the \( r^{th} \) order statistic of
a sample of \( n \) values of \( d \) from the posterior distribution and again let \( F \) be the reference
distribution of \( d \). Using (3.16) then we get

\[
P(D(r) > t) = 1 - F_{D(r)}(t) \\
\leq 1 - \max \left\{ 0, \frac{nF(t) - r + 1}{n - r + 1} \right\}.
\]

(3.18)

The right side of (3.18) can be shown to be equal to \( \min \left\{ 1, \frac{n(1-F(t))}{n-r+1} \right\} \) by verifying that for
real numbers \( a \) and \( b \) then \( 1 - \max\{a, b\} = \min\{1 - a, 1 - b\} \), and noting that \( 1 - \frac{nF(t)-r+1}{n-r+1} \)}
\[ P(D(r) > t) \leq \min \left\{ 1, \frac{n(1 - F(t))}{n - r + 1} \right\} , \quad (3.19) \]

which gives an upper bound on the probability \( P(D(r) > t) \) in the left side of (3.17).

To apply these bounds as a model checking method, \cite{Yuan:2011} propose, for a given sample \( \{d(y, \tilde{\theta})\} \) from the posterior of \( d \), to compute the upper bound in the right side of (3.19) for each of the order statistics. In this, for each \( r = 1, \ldots, n \) we let \( t = d(r) \), the \( r^{th} \) order statistic of the sample, and calculate the \( n \) different upper bounds on the right side of (3.19). The overall minimum of the resulting probabilities, denoted \( p_{\min} \), is looked at to determine evidence of model lack-of-fit.

To illustrate how this is useful, first remember that our null hypothesis is that the model fits the data, and so under this \( d(y, \tilde{\theta}) \) will follow the reference distribution \( F \). This also implies that for a dependent sample \( \{d(y, \tilde{\theta})\} \) of size \( n \) from the posterior of \( d \), the order statistics should be similar to the order statistics from a dependent sample from \( F \). If an order statistic value from the sample is extreme relative to the reference distribution, then the probability in the left of (3.19) will be small.

For example, consider the posterior sample of 1,000 values of the PDM \( d_1 = \sum_{i=1}^{10} (y_i - \theta_i)^2 / \sigma^2 \) that is compared to the \( \chi^2_{10} \) density in the bottom of figure 3.8. For this sample the \( r = 700^{th} \) order statistic was \( d_{1(700)} = 26.2 \). The \( \chi^2_{10} \) distribution function evaluated at 26.2 is 0.997. Applying (3.19) we see that if we were to draw a sample of 1,000 posterior values of \( d_1 \) (computed from data arising under the model), then the probability of observing a 700\(^{th}\) order statistic value \( d_{1(700)} \) as extreme as 26.2 is

\[ P(d_{1(700)} > 26.2) \leq \min \left\{ 1, \frac{1,000(1 - 0.997)}{1,000 - 700 + 1} \right\} \approx 0.01. \quad (3.20) \]

So, the probability of observing a 700\(^{th}\) order statistic value of the measure \( d \) as extreme as 26.2 is approximately at most 0.01. This is evidence that the data that produced the
posterior sample of $d_1$ values was not generated under the model. As mentioned above, Yuan and Johnson (2011) recommend computing this value for all $r = 1, \ldots, 1,000$ and taking the overall minimum of the values.

The true probability $P(D(r) > t)$ in the left side of (3.19) is very difficult to compute since the exact distribution of order statistics of identically distributed dependent samples is difficult to compute (see David and Nagaraja 1970). Using just an upper bound means that this method may tend to be more conservative than having the true probability. Taking this into consideration, Yuan and Johnson (2011) give different criterion for how to evaluate the resulting value of $p_{\text{min}}$.

Yuan and Johnson (2011) propose the following criteria for using the value of $p_{\text{min}}$ to evaluate model adequacy. They term this criteria the rule of thumb, and we will refer to this by this name.

1. If $p_{\text{min}} < 0.05$, there is strong evidence of model lack-of-fit.

2. If $0.05 \leq p_{\text{min}} < 0.25$, there is some evidence of inadequacy, and the posterior distribution of PDMs should be evaluated more precisely, possibly using the prior predictive posterior simulation (PPPS) based method of Dey et al. (1998), or other PDMs should be considered.

3. If $p_{\text{min}} \geq 0.25$, the PDM does not provide evidence of inadequacy.

We note that the bound given in (3.19) only considers having large values of PDMs as indicating lack of fit. In fact, when Yuan and Johnson (2011) applied this they mainly looked at PDMs based on squared residuals so one would naturally look for large values in this case. However, what if one is using a PDM in which small values could indicate lack of fit? For example, look at our PDM $d_{1i}(y_i, \tilde{\theta}_i) = (y_i - \tilde{\theta}_i)/\sigma$ for each $i = 1, \ldots, 10$ which has a $N(0,1)$ reference distribution. In this case, small or large values of the residual could indicate lack-of-fit. So, what is needed is to incorporate a bound on the order statistics that quantifies extremeness of small values.
To answer this, we look back at the paper by Caraux and Gascuel (1992) and see that there is indeed such a bound. This is given in the following proposition,

**Proposition 1.** Let \( X_1, \ldots, X_n \) be a set of \( n \) identically distributed random variables (with c.d.f. \( F \)), then

\[
F_{X(r)}(x) \leq \inf \left( \frac{nF(x)}{r}, 1 \right) \quad (3.21)
\]

This would imply

\[
P(D(r) \leq t) \leq \min \left\{ \frac{nF(t)}{r}, 1 \right\}. \quad (3.22)
\]

In the same fashion as before, for each \( t = d(r) \) we can compute the right side of (3.22) to obtain an upper bound for the left side of (3.22). In the following chapters we will apply these two bounds when evaluating model fit when using PDMs. Since we will use PDMs like \( d_i(y_i, \hat{\theta}_i) = (y_i - \hat{\theta}_i)/\sigma \) which can show lack of fit with small or large values, we will compute the overall minimum between the two bounds in (3.19) and (3.22), and take the minimum between these two to compute a new value of \( p_{\text{min}} \). That is,

\[
p_{\text{min}} = \min \left\{ \min_{r=1} \left\{ 1, \frac{n(1 - F(t))}{n - r - 1} \right\}^{*}, \min_{r=1} \left\{ \frac{nF(t)}{r}, 1 \right\}^{*} \right\}. \quad (3.23)
\]

The rule of thumb mentioned above will then be applied to determine rejection.

Johnson (2007) gave the following things to consider with identifying \( p_{\text{min}} \): “...if the distribution of the pivotal quantity is not exact, extreme values of the \( r/n \) quantile should be regarded with caution since in that case the marginal distribution of the pivotal quantity in the extreme tails of the distribution will also not be exact.” This last point is meant to apply to discrepancy measures that are pivotal asymptotically. That is, they only asymptotically have a reference distribution that is free of unknown parameters. In this case Yuan and Johnson (2011) recommend discarding a certain percentage of the top and bottom ordered values of the posterior sample of a PDM before computing the value of \( p_{\text{min}} \).
3.5 Applications to IRT Models: A Preview

In the next chapter we will discuss how these model checking procedures can be applied to Bayesian IRT models. In the examples and discussions in sections 3.3.3, 3.4.4.2, and 3.4.4.3 we saw how the PPPS and PDM methods can be used to identify outlying data values or assess the fit of a whole data set under a given model. We will be using these for a similar application in IRT models. Specifically, we will investigate how these procedures can be used to examine how well an examinee’s vector of responses $y_i = (y_{i1}, \ldots, y_{iK})'$, $i = 1, \ldots, I$, fit with the specified IRT model. Our goal will be to identify examinee responses that are outlying, or aberrant, under the model. Several statistics used for this in a classical setting will be introduced and we will discuss how they can be used in a Bayesian IRT setting. We will also be introducing how the latent variables $Z_{ik}$ discussed in chapter 1 can be used as model checking measures for this as well.

The specific methods we will be using are the posterior predictive, PPPS and PDM model checking methods. Much of the research done in Bayesian IRT model checking has employed the posterior predictive method (i.e. Sinharay and Johnson 2003, Sinharay 2005, Glas and Meijer 2003, Sinharray et al. 2006). Some research has been done using the prior predictive method (see Zhang 2008), but due to its prevalent use we choose to use the posterior predictive method. Since this method is used more often, this will give us a kind of baseline for which to compare the two new methods we have introduced. As we have mentioned, the PPPS and PDM methods are relatively new to Bayesian model checking, and to the writer’s knowledge this is the first time they will be applied to model checking in Bayesian IRT models.

After our initial discussion, the objective of the next chapter will be to conduct a simulation study comparing the different discrepancy measures used under the posterior predictive, PPPS, and PDM methods. As mentioned, most research investigating examinee response fit (Glas and Meijer 2003) has been done using the posterior predictive method, employing
many of the fit measures that we will introduce. So, we have only been able to see which measures work best under this method so far. In our simulation study we can get an idea of the performance of these different measures under the PPPS and PDM methods.
CHAPTER 4
CHECKING PERSON FIT IN BAYESIAN 2PN IRT MODELS

4.1 Introduction

In this chapter we will investigate how the posterior predictive (PP), prior predictive posterior simulation (PPPS), and pivotal discrepancy measure (PDM) model checking methods can be applied to Bayesian IRT models. As mentioned at the end of chapter 3 the application of these methods will be to identify any examinee(s) response vectors that is outlying under the assumed IRT model.

The outline of the chapter will be as follows. Section 4.2 will discuss some common types of misfit in IRT models and go more in depth on the type of misfit we will focus on as part of our research. Sections 4.3 and 4.4, respectively, will introduce a series of classical IRT model checking statistics and show an example of how these statistics are used. In section 4.5, we will discuss how the Bayesian model checking methods (PP, PPPS, PDM) can be used along with the statistics of section three. Also, we will introduce how the latent variables $Z_{ik}$ (see chapter 1) can be used as a model checking measure for Bayesian IRT models.

Section 4.6 will show an example illustrating how the different model checking statistics will be used under each of the three Bayesian methods. Lastly, in section 4.7 we will present the results of a simulation study that was conducted to investigate the performance of the different model checking statistics under the PP, PPPS, and PDM model checking methods when applied to data generated under a Bayesian 2PN IRT model. We will look at the type I error rates of the of the respective measures. Also, we will look at the detection rates of the statistics in identifying examinees’ response patterns that have been simulated under four specific types of model violation.
The Bayesian IRT model that we will be focusing on in this chapter is the Bayesian 2PN IRT model, specifically that which was specified in section 2.4.3.2. We will be using it enough in this chapter that it may be worth re-stating it here. The model is given by,

\[
y_{ik} = \begin{cases} 
1, & \text{if } Z_{ik} > 0 \\
0, & \text{if } Z_{ik} \leq 0
\end{cases} \tag{4.1}
\]

\[Z_{ik} \sim N(a_k \theta_i - b_k, 1) \tag{4.2}\]

\[\theta_i \sim N(0, 1), \tag{4.3}\]

\[a_k \sim TN(0, \infty)(1,.50), \tag{4.4}\]

\[b_k \sim N(0, 1), \tag{4.5}\]

for \(i = 1, \ldots, I\) and \(k = 1, \ldots, K\). In this model, we include the latent variable \(Z\) (which we will discuss in much more detail in section 4.5.1.2) with a normal prior distribution with mean \(a_k \theta_i - b_k\) and unit variance. The examinee abilities are given a \(N(0, 1)\) prior distribution. The item discrimination and difficulty parameters were given independent priors of \(TN(0,\infty)(1,.50)\) and \(N(0,1)\), respectively.

### 4.2 Discussion of Aspects of Fit in IRT Models

Any statistical model requires a set of assumptions about the data to which it applies. Two assumptions that are commonly checked in IRT models are dimensionality and local independence (LI). As in analysis done with other models, identifying data values that are potentially outlying under the model is also of concern. Finally, checking the overall goodness-of-fit of an IRT model is of interest.

Dimensionality in IRT models refers to the number of latent abilities that are thought to underlie examinee performance in answering a set of test items. Unidimensional IRT models assume that only one latent trait is needed to account for examinee performance
on a test. Multidimensional IRT models assume that more than one latent trait is needed. For example, a math test where items are requiring just computation may only need an examinee’s mathematics ability, but if a test has many word problems it might be that performance would also rely on an ability for reading comprehension.

Local independence was discussed in chapter 1 and is the assumption that, conditional on the ability and item parameters, an examinee’s responses to all test items are statistically independent. This is a common assumption in 1P, 2P, and 3P IRT models. There are some models that do not make the this assumption, which we will study in the next chapter. Checking the LI assumption is important because, as we will see in the next chapter, fitting an IRT model with this assumption violated by a set of data can lead to biased parameter estimates.

In IRT models there is a concept of outlying data values under the model. Instead of single response data values, $y_{ik}$ we are more concerned with how well all the examinee responses for a given item fit the model. This is referred to as checking item fit. This can be used to perhaps: (1) determine if certain items do net meet the proper dimensionality assumption of the model (i.e. if the model assumes unidimensionality and an item tests more than one ability) and (2) to identify faulty items, i.e. if an item has all examinees answering it incorrectly, maybe it is a poorly written question.

In IRT models we are also concerned with assessing the fit of each examinee’s item responses to a given model. This is referred to as person fit and is used to identify examinees’ aberrant response patterns. Finally, we may also want to assess the overall model goodness-of-fit to a set of data. For example, if we choose to fit a 1P IRT model to a set of response data, then we are making the assumption that the discrimination parameter for all items is one. We may wish to check if this particular model is a good fit, versus say a 2P IRT model.

We will focus on these later two areas of model assessment. In this chapter we will look at methods of assessing person fit in Bayesian 2PN IRT models. In the next chapter, we will look at procedures for assessing the goodness-of-fit of a Bayesian 2PN IRT Model. We will
begin now with a more in depth explanation of person fit.

4.2.1 Explanation & Background of Person Fit

When taking a test, the binary responses of an examinee are recorded in a $K \times 1$ vector referred to as an *item response pattern* or just *response pattern*. Examining *person fit* or *examinee fit* is essentially looking at how well an examinee’s item response pattern fits the given IRT model. Under the assumptions of IRT models, in particular that the probability of an examinee getting a given item correct is an increasing function of their ability, means that response patterns should behave as Guttman patterns (Guttman 1950). That is, if all items were ordered from easiest to hardest, then the pattern of responses would have mostly all of the incorrect answers at the end of the pattern.

For example, in table 4.1 pattern 1 and 2 follow Guttman patterns. Whatever an examinee’s ability, the examinee will get the easier items correct, but eventually the difficulty is too great and the examinee will have incorrect responses. Patterns 3 and 4 are examples of violations of this expected pattern. We see that these examinees got mostly the easy questions wrong while getting the more difficult questions correct.

<table>
<thead>
<tr>
<th>Pattern #</th>
<th>Response Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1111100000</td>
</tr>
<tr>
<td>2</td>
<td>1111010000</td>
</tr>
<tr>
<td>3</td>
<td>0100101101</td>
</tr>
<tr>
<td>4</td>
<td>0000011111</td>
</tr>
</tbody>
</table>

Table 4.1 Examples of different response patterns.

Different reasons have been postulated in the literature as to how examinees can produce unlikely patterns under a given model. For example, an examinee may have prior knowledge of answers to some of the more difficult questions, or perhaps he or she is just guessing his or her way through the test. Examinee ability estimation in IRT models is relative, and dependent on the performance of other examinees taking the test. Due to this, it is
important to flag examinees whose response patterns are not in accordance with the model. In this situation, researchers may choose to eliminate such examinees from the examinee pool and re-fit the model. Person fit statistics are used to determine the fit of an examinee’s response pattern with the given IRT model.

4.3 Classical Person Fit Measures

Classical person fit statistics can be divided into many different groups. Meijer and Sijtsma (2001) give a very good review of the most commonly used person fit statistics, and Karabatsos (2003) performs a simulation study comparing thirty-six different parametric and non-parametric person fit measures. Other work such as that in Nering and Meijer (1998) and Glas and Meijer (2003) also compare some of these statistics in smaller groups.

There are three specific types that we will look at. The following person fit statistics are among the most widely used in the literature that the author has studied. The first are measures that are based on a likelihood function. The second are statistics that are based on residuals of the form $y_{ik} - E(y_{ik})$ that reflect the discrepancy between expected item scores under the model and observed item scores. For each $i$ and $k$, $y_{ik}$ is a Bernoulli random variable and so its expected value $E(y_{ik}) = p_{ik}$, the probability of success. In the statistics we introduce below, as a notational convenience, we will use $p_{ik}$ instead of $E(y_{ik})$. The third type of person fit statistic we will look at is in the category of “caution indices” (Sato 1975). These measures involve computing ratios of covariances between certain parts of a response matrix $Y$. It is difficult to describe these measures further in just a few lines, so we will hold on this until later.

We will introduce the different classical person fit measures. We give them in their theoretical forms involving $p_{ik}$. In practice, under an assumed IRT model $p_{ik}$ is estimated using the MLE’s of the item and ability parameters. For example, under a 2PN model the probability of success $p_{ik} = \Phi(a_k \theta_i - b_k)$ would be estimated as $\hat{p}_{ik} = \Phi(\hat{a}_k \hat{\theta}_i - \hat{b}_k)$ where $\hat{a}_k$ and $\hat{b}_k$ are the MLE’s of the item discrimination and difficulty parameters, respectively, and
\( \hat{\theta}_i \) is the MLE of the examinee’s ability. Later on in section 4.4, we will show an example of how each measure is computed and used in practice.

### 4.3.1 Classical Likelihood Based Measures

#### 4.3.1.1 Levine and Rubin (1979)

As a measure of person fit, Levine and Rubin (1979) proposed the measure \( L_0 \) which is the log-likelihood function summed over items \( k = 1, \ldots, K \) for each examinee \( i = 1, \ldots, I \).

The discrepancy measure is given as,

\[
L_{0,i} = \sum_{k=1}^{K} [y_{ik}\log(p_{ik}) + (1 - y_{ik})\log(1 - p_{ik})],
\]

(4.6)

where \( p_{ik} = \Phi(a_k\theta_i - b_k) \). The nice thing about this measure is that it is simple to compute, but as Drasgow et al. (1985) showed, “…the conditional distribution of \( L_0 \) given \( \hat{\theta} \) varies as a function of \( \hat{\theta} \) and it is difficult to interpret the magnitude of \( L_0 \) directly”. We do not investigate this measure here, but mention it to introduce the next measure which we will use. This is a standardized version of \( L_0 \), given by Drasgow et al. (1985) which overcomes this issue of dependence on the value of \( \theta \).

#### 4.3.1.2 Standardized \( L_z \) of Drasgow, Levine & Williams (1985)

Drasgow et al. (1985) proposed a standardization of the log-likelihood statistic, \( L_0 \). This person fit measure, denoted \( L_z \), is given by

\[
L_{z,i} = \frac{L_{0,i} - E(L_{0,i})}{\sqrt{Var(L_{0,i})}}, \quad i = 1, \ldots, I,
\]

(4.7)
where $E(L_{0,i})$ and $Var(L_{0,i})$ are the expected value and variance of $L_{0,i}$ respectively. These are given by,

$$E(L_{0,i}) = \sum_{k=1}^{K} [p_{ik}\log(p_{ik}) + (1 - p_{ik})\log(1 - p_{ik})]$$

(4.8)

and

$$Var(L_{0,i}) = \sum_{k=1}^{K} p_{ik}[1 - p_{ik}][\log\frac{p_{ik}}{1 - p_{ik}}]^2.$$  

(4.9)

Drasgow et al. (1985) purported that $L_z$ was asymptotically standard normally distributed. In practice, it is with reference to this distribution that the statistic $L_z$ is compared. However, as mentioned above, the item and ability parameters are estimated with their respective MLE’s. It has been suggested that because of this the null distribution of $L_z$ is not very close to standard normal, but instead that under the true values of the parameters the null distribution of $L_z$ is asymptotically standard normal (Molenaar and Hoijttink 1990, Armstrong et al. 2007).

4.3.2 Classical Residual Based Measures

The following measures are based on residuals of the observed score minus the expected score. These can be divided into two groups: total person fit statistics and between person fit statistics. Total fit statistics look at summing squares of residuals $y_{ik} - p_{ik}$ over all items $K$. Two that we will look at will be the unweighted and weighted mean of squared residuals.

Between fit statistics look at summing residuals over disjoint subsets of test items (Smith 1982, 1985). As described in Smith (1985) “...if the data fit the model, the overall ability estimate should accurately predict the person’s score on any subset of items. By comparing the person’s predicted score with observed score on any subset of items it is possible to test the fit of the data to the model.” For example, suppose we rank all items in order of difficulty and split them into two groups with the 50% easiest items in one group and the 50% hardest items in another. If an examinee’s responses follow according to the model, there should not be any large discrepancy between the scores and expected scores between each group.
Each of the two types of residual based statistics plays a different role in detecting response aberrancies. The total fit statistics are good at detecting systematic types of response violations such as _random responding_ - where a student might haphazardly answer items all through a test. The between fit statistics are better at detecting violations that occur in particular portions of items such as _guessing_ - where students answer most items according to their ability, but if the difficulty level of a particular item(s) is too high they guess at the answer instead of answering it honestly. We will now discuss these statistics and how they are computed in practice.

### 4.3.2.1 Unweighted Mean Square Fit Statistic by Wright and Stone (1979)

Wright and Stone (1979) proposed a person fit statistic based on squared standardized residuals. For a given examinee $i$, with responses to the $K$ items, $y_{i1}, \ldots, y_{iK}$ and corresponding probabilities of correctly answering the items $p_{i1}, \ldots, p_{iK}$, then Wright and Stone (1979) look at the standardized residuals,

$$z_{ik} = \frac{y_{ik} - p_{ik}}{\sqrt{v_{ik}}}, \quad (4.10)$$

for $k = 1, \ldots, K$ where $v_{ik} = p_{ik}(1 - p_{ik})$ is the variance of $y_{ik}$. Recall that conditional on the item and ability parameters, the responses $y_{i1}, \ldots, y_{iK}$ are assumed to be locally independent and form a set of $K$ independent Bernoulli random variables with expected value $E(y_{ik}) = p_{ik}$ and variance $\text{Var}(y_{ik}) = p_{ik}(1 - p_{ik})$. This means that the random variables $z_{i1}, \ldots, z_{iK}$ are standardized in that they have mean zero and variance one.

From here, a “chi-squared” like statistic (Smith 1982) is created by squaring the $z_{ik}$’s and summing them over the items. That is,

$$\chi^2_{K,i} = \sum_{k=1}^{K} z_{ik}^2 = \sum_{k=1}^{K} \frac{(y_{ik} - p_{ik})^2}{v_{ik}}, \quad (4.11)$$
The statistic used by Wright and Stone (1979) is,

\[ U_i = \frac{1}{K} \sum_{k=1}^{K} z_{ik}^2, \]

(4.12)

which is termed in Wright (1980) as an unweighted mean square statistic. We will discuss the weighted version of this mean square statistic shortly.

The reason why this is referred to as a “chi-squared” like statistic is that (4.11) would be a chi-squared random variable with \( K \) degrees of freedom if \( z_{ik} \sim N(0, 1) \). The standardized residuals could be claimed to be approximately standard normal if the \( y_{ik} \) were binomial random variables and the number of trials, \( n \), was sufficiently large. However, the \( y_{ik} \) are Bernoulli random variables and so \( n = 1 \) and there is no real approximation to be made here. Nonetheless, Wright and Stone (1979) and Wright (1980) continue as if \( z_{ik} \) were approximately standard normally distributed and use some results based on this to analyze \( U_i \).

Wright (1980) uses a cube-root transformation introduced by Wilson and Hilferty (1931) that transforms a \( \chi^2_N \) random variable into an approximate standard normal random variable. This transformation is given by,

\[ t(V) = (V^{1/3} - 1) \frac{3}{S} + \frac{S}{3}, \]

(4.13)

where \( V \) is a \( \chi^2_N/N \) random variable and \( S \) is the standard deviation of \( V \). Wright (1980) show the standard deviation of \( U_i \) to be,

\[ S_{U_i} = \left[ \sum_{k=1}^{K} v_{ik}^{-1} - 4K \right] \cdot K^{-1}. \]

(4.14)

With this, if (4.11) held as a \( \chi^2_K \) random variable then \( U_i \) would be a \( \chi^2_K/K \) random variable and could be transformed by \( t \) in (4.13) to a standard normal random variable.

Wright (1980) acknowledge that the sum of squared residuals in (4.11) is not a true
chi-squared random variable and so claiming the distribution of $t(U_i)$ to be $N(0, 1)$ is not theoretically valid. Wright (1980) also note that in practice, the item and ability parameters are unknown and so $p_{ik}$ must be estimated using item and ability parameter estimates. This results in not knowing the exact distributional properties of $t$ in (4.13). However, Mead and Wright (1980) and Smith (1982) have indicated through simulation that the distribution of $t(U_i)$, where $U_i$ is evaluated using estimated item and ability parameters, is close to standard normal when the data fit the model.

### 4.3.2.2 Weighted Mean square Fit Statistic by Wright (1980)

An issue with the statistic $U$ is that it can be sensitive to outlying responses. Wright (1980) considered an alternative to the mean sum of squared residuals, $U$, and that is a weighted sum of squared residuals. This is given by,

$$W_i = \frac{z_{i1}^2 v_{i1} + z_{i2}^2 v_{i2} + \ldots + z_{iK}^2 v_{iK}}{v_{i1} + v_{i2} + \ldots + v_{iK}}$$

$$= \frac{\sum_{k=1}^{K} z_{ik}^2 v_{ik}}{\sum_{k=1}^{K} v_{ik}}$$

$$= \frac{\sum_{k=1}^{K} (y_{ik} - p_{ik})^2}{\sum_{k=1}^{K} v_{ik}}, \quad (4.15)$$

for $i = 1, \ldots, I$. In this, each squared residual $z_{ik}^2$ is weighted by the variance $v_{ik}$. Since the variance is smallest for probability values $p_{ik}$ closest to zero or one, then the contribution to $W_i$ by an outlying response is reduced.

Smith (1982) computes the standard deviation of $W_i$ to be,

$$S_{W_i} = \frac{\left[ \sum_{k=1}^{K} v_{ik} - 4 \sum_{k=1}^{K} v_{ik}^2 \right]^{1/2}}{\sum_{k=1}^{K} v_{ik}}. \quad (4.16)$$

Smith (1982) applied the cube-root transformation $t$ in (4.13) to $W_i$ and found that the distribution of $t(W_i)$ was close to standard normal when the data were generated under the
Table 4.2 Simulated probabilities of a correct response to 5 items for a given examinee. The variance $v_{ik} = p_{ik}(1 - p_{ik})$ is also computed for each item.

As an example to illustrate the difference between the statistics $U$ and $W$, suppose we have have a short test with five items. For a hypothetical examinee, we have simulated five probabilities of correctly answering each item and the corresponding responses to each item. These are given in Table 4.2 along with the calculated variances for each item. We calculate the statistics $U$ and $W$ as,

\[
U = \frac{1}{5} \sum_{k=1}^{5} \frac{(y_{ik} - p_{ik})^2}{v_{ik}}
\]
\[
= \frac{0.236 + 0.037 + 0.266 + 0.098 + 0.238}{5}
\]
\[
= 0.184, \quad (4.17)
\]

and

\[
W = \frac{\sum_{k=1}^{5} z_{ik}^2 v_{ik}}{\sum_{k=1}^{5} v_{ik}}
\]
\[
= \frac{0.236(0.155) + 0.037(0.035) + 0.266(0.167) + 0.098(0.082) + 0.238(0.172)}{0.602}
\]
\[
= 0.228. \quad (4.18)
\]
To compute the cube-root transformation we compute the standard deviation of each statistic,

\[ S_U = \left[ \sum_{k=1}^{5} v_{ik}^{-1} - 4(5) \right]^{1/2} / 5 \]
\[ = 1.258, \quad (4.19) \]

and

\[ S_W = \left[ \sum_{k=1}^{5} v_{ik} - 4 \sum_{k=1}^{5} v_{ik}^{2} \right]^{1/2} / \sum_{k=1}^{5} v_{ik} \]
\[ = 0.826, \quad (4.20) \]

and get,

\[ t(U) = (0.184^{1/3} - 1) (3/1.258) + (1.258/3) \]
\[ = -0.609, \quad (4.21) \]

and

\[ t(W) = (0.228^{1/3} - 1) (3/0.826) + (0.826/3) \]
\[ = -1.137. \quad (4.22) \]

Now look at say, item 2, where the probability of success is 0.036 and the item was answered incorrectly. Suppose the item was gotten correct instead, by accident or lucky guess. Then the new response pattern would be: 0 1 1 0 1. We re-calculate \( U \) and \( W \) to see
the effect of this outlying response.

\[
U = \frac{1}{5} \sum_{k=1}^{5} \frac{(y_{ik} - p_{ik})^2}{v_{ik}} \\
= [0.236 + 26.938 + 0.266 + 0.098 + 0.238]/5 \\
= 5.564,
\]

(4.23)

and

\[
W = \frac{\sum_{k=1}^{5} z_{ik}^2 v_{ik}}{\sum_{k=1}^{5} v_{ik}} \\
= [0.236(0.155) + 26.938(0.035) + \\
0.266(0.167) + 0.098(0.082) + \\
0.238(0.172)]/0.602 \\
= 1.750.
\]

(4.24)

We can see how the residual corresponding to the outlying response, 26.938, is much larger than before, 0.037. In the computation of \( W \), this value is scaled by the variance of 0.035, so its effect is much less than in \( U \).

We compute the new transformed values of \( U \) and \( W \) and get,

\[ t(U) = t(5.564) = 2.261, \]

(4.25)

and

\[ t(W) = t(1.750) = 1.022. \]

(4.26)

Comparing these values to a standard normal scale, we see the value of \( t(U) \) is extreme and \( t(W) \) is not. We recognize that these are only five items, and the purpose of these statistics is to identify aberrant response patterns. However, it is more useful if the statistic is not too affected by a few outlying responses.
4.3.2.3 Unweighted Between-Set Fit Statistic by Smith (1985)

Mead (1980) first introduced a between-set fit statistic which looks at summing squared residuals, as with $W$ and $U$, but over disjoint subsets of items instead of over all items. Smith (1982, 1985) then investigated this under the Rasch (or 1PL) IRT model. This statistic has a weighted and unweighted form. In this work we will focus on the unweighted version which we denote as $UB$ (Smith 1985).

To use this person fit measure, it is assumed that the test can be divided into $S$ non-overlapping subsets which we denote $A_s$, $s = 1, \ldots, S$. The statistic $UB$ is given by,

$$UB_i = \frac{1}{S - 1} \sum_{s=1}^{S} \left[ \frac{\sum_{j \in A_s} (y_{ij} - p_{ij})}{\sum_{j \in A_s} v_{ij}} \right]^2,$$

for $i = 1, \ldots, I$. We can see that $UB_i$ is essentially measuring the discrepancy between the observed sum of scores $\sum_{j \in A_s} y_{ij}$ and its expected value $\sum_{j \in A_s} p_{ij}$ in subset $A_s$. The term

$$\frac{\sum_{j \in A_s} (y_{ij} - p_{ij})}{\sum_{j \in A_s} v_{ij}} = \frac{\sum_{j \in A_s} y_{ij} - \sum_{j \in A_s} p_{ij}}{\sum_{j \in A_s} v_{ij}},$$

is a standardized residual for $\sum_{j \in A_s} y_{ij}$ since $Var(\sum_{j \in A_s} y_{ij}) = \sum_{j \in A_s} Var(y_{ij}) = \sum_{j \in A_s} v_{ij}$, under the local independence assumption. So, $UB$ is a sum of squared residuals “like” statistic such as $U$ or $W$.

Squaring the summed residuals over subsets in this statistic posses a possible advantage in detecting response violation types where the majority of items with aberrant responses may fall into one of the subsets. For example, suppose someone who engages in guessing is one of lower ability, but tends to get correct answers to the most difficult items. Having a lower ability implies that the success probabilities will tend to be smaller, especially for very difficult items. Suppose we were to split the items into subsets based on item difficulty. Then the subset made of the most difficult items would yield very high residuals, because the examinee’s responses would be 1’s and their success probabilities would be small. Squaring
the summed residuals over this subset would yield a very large value, making the overall statistic value larger.

Smith (1982) investigates applying the cube-root transformation \( t \) in (4.13) to \( UB \). The standard deviation of \( UB \) was found to be,

\[
S_{UB} = \left( \frac{2}{S - 1} \right)^{1/2}.
\] (4.28)

Using simulated data under the Rasch model, evidence is found that the distribution of \( t(UB) \) is close to a standard normal distribution. Exactly how close the computed sample values of \( t(UB) \) were to being standard normal was not mentioned. The main details that were given were that the distribution of sample values seemed to be symmetric and that approximately only 2.5% of the values were greater than 2. Due to this, Smith (1985) uses 2 as a critical value to judge the extremeness of a computed value of \( t(UB) \).

4.3.3 Classical Standardized Extended Caution Indices of Tat-suoka (1984)

Sato (1975) proposed a person fit measure called the Caution Index which is defined as,

\[
C_i = 1 - \frac{\text{cov}(y_i, y,)}{\text{cov}(x_i, y,)}.
\] (4.29)

The second term in (4.29) is the ratio between two covariances. The numerator is the covariance between the response vector \( y_i = (y_{i1}, \ldots, y_{iK})^t \) of examinee \( i \) and the column-sum vector of item responses \( y, = (\sum_{i=1}^I y_{i1}, \ldots, \sum_{i=1}^I y_{iK})^t \). The denominator is the covariance between a Guttman vector \( x_i \), with the same total scores as the response vector \( y_i \), and the column-sum vector. In particular, \( x_i \) is formed by taking the total score \( s_i = \sum_{k=1}^K y_{ik} \) of response vector \( y_i \) and forming a Guttman vector whose first \( s_i \) elements are 1's and the remaining elements are 0's. What this ratio does is to “...compare the degrees of
similarity between the observed vector \( y_i \) and its Guttman scale vector \( x_i \), respectively with the column sum vector of correct answers for the items” (Tatsuoka 1984). The Guttman vector \( x_i \) representing the “ideal” response vector for examinee \( i \) under the model.

Tatsuoka and Linn (1983) introduced a set of six Extended Caution Indices which were extensions of the Caution Index of (4.29). The later research of this chapter will incorporate measures based on the second and fourth of these extended indices, which will be explained later. We show these two extended indices are here,

\[
ECI_{2i} = 1 - \frac{\text{cov}(y_i, G)}{\text{cov}(\hat{p}_i, G)} \tag{4.30}
\]

\[
ECI_{4i} = 1 - \frac{\text{cov}(y_i, \hat{p}_i)}{\text{cov}(G, \hat{p}_i)} \tag{4.31}
\]

where \( \hat{p}_i = (\hat{p}_{i1}, \ldots, \hat{p}_{iK})^t \) is the vector of estimated probabilities of success for examinee \( i \), and \( G \) is the column mean vector of the estimated \( I \times K \) response probability matrix \((\hat{p}_{ik})\). That is, the vector containing the average estimated probability of a correct response to each item, across all examinees.

The extended caution index \( ECI_{2i} \) replaces the Guttman vector \( x_i \) in (4.29) by the estimated probability vector \( \hat{p}_i \) and the column sum vector \( y_i \) by \( G \). The extended caution index \( ECI_{4i} \) replaces \( x_i \) by \( G \) and so the ratio of covariances in (4.31) compares the similarity of the relationship between \( y_i \) and \( \hat{p}_i \) and that of \( G \) and \( \hat{p}_i \).

Tatsuoka (1984) then introduced the Standardized Extended Caution Indices which were standardizations of the extended caution indices introduced in Tatsuoka and Linn (1983). The standardizations of \( ECI_{2i} \) and \( ECI_{4i} \) are given here as,

\[
ECI_{2z,i} = \frac{ECI_{2i} - E(ECI_{2i})}{\sqrt{Var(ECI_{2i})}} \tag{4.32}
\]
\[ ECI_{4_{z,i}} = \frac{ECI_{4_{i}} - E(ECI_{4_{i}})}{\sqrt{Var(ECI_{4_{i}})}}. \] (4.33)

One of the reasons for standardizing the original ECI’s as explained in Tatsuoka (1984), was to correct for the fact that the original ECI’s gave inflated values at extreme values of \( \hat{\theta}_i \). Tatsuoka (1984) also explain that after investigating the relationship between all of the standardized ECI’s, that ECI_{2_{z,i}} and ECI_{4_{z,i}} correlate very highly will the others. Due to this they choose to conduct their research on ECI_{2_{z,i}} and ECI_{4_{z,i}} only, as person fit measures, which they henceforth denote as \( \xi_{1,i} \) and \( \xi_{2,i} \), respectively.

The derivations of the expectations and variances of ECI_{2_{z,i}} and ECI_{4_{z,i}} are shown in Tatsuoka (1984). It is also shown that (4.32) and (4.33) can be expressed in the equations,

\[
\xi_{1,i} = \frac{\sum_{k=1}^{K} p_{ik} (G_k - G)}{\left( \sum_{k=1}^{K} \sigma_{ik}^2 (G_k - G)^2 \right)^{1/2}} - \frac{\sum_{k=1}^{K} y_{ik} (G_k - G)}{\left( \sum_{k=1}^{K} \sigma_{ik}^2 (G_k - G)^2 \right)^{1/2}},
\]

(4.34)

and

\[
\xi_{2,i} = \frac{\sum_{k=1}^{K} p_{ik} (p_{ik} - T_i)}{\left( \sum_{k=1}^{K} \sigma_{ik}^2 (p_{ik} - T_i)^2 \right)^{1/2}} - \frac{\sum_{k=1}^{K} y_{ik} (p_{ik} - T_i)}{\left( \sum_{k=1}^{K} \sigma_{ik}^2 (p_{ik} - T_i)^2 \right)^{1/2}},
\]

(4.36)

where \( G_k = \frac{1}{I} \sum_{i=1}^{I} p_{ik} \) is the \( k^{th} \) column mean of the probability matrix \( (p_{ik}) \) and \( G = \frac{1}{K} \sum_{k=1}^{K} G_k \) is the overall mean of \( (p_{ik}) \). The term \( T_i = \frac{1}{K} \sum_{k=1}^{K} p_{ik} \) is the \( i^{th} \) row mean of \( (p_{ik}) \) and \( \sigma_{ik}^2 = p_{ik}(1 - p_{ik}) \) is the variance of the response \( y_{ik} \).
The measures $\xi_1$ and $\xi_2$ will be sensitive to item score patterns with incorrect responses to easy items and correct responses to difficult ones (Glas and Meijer 2003). Both measures will be positive when this happens. The more a response pattern deviates from a Guttman pattern the larger the values of $\xi_1$ and $\xi_2$ will be. The more closely a response pattern follows a Guttman pattern the smaller the value of $\xi_1$ and $\xi_2$ will be.

See Tatsuoka (1984) for a full explanation, but we can briefly describe this here. To see why $\xi_1$ is negative for Guttman type response patterns, consider arranging the column means $G_k$, $k = 1, \ldots, K$, in descending order. That is,

$$G_{(K)} \geq \ldots \geq G_{(2)} \geq G_{(1)} \geq 0,$$

(4.38)

where $G_{(j)}$ is the $j^{th}$ ordered value of $G_k$, $k = 1, \ldots, K$. This implies then,

$$G_{(K)} - G \geq \ldots \geq G_{(2)} - G \geq G_{(1)} - G.$$

(4.39)

Since $G$ is the average of the $G_k$, $k = 1, \ldots, K$, then for some $j$, $1 \leq j \leq K$, we have $G_{(k)} - G$ positive for $k \geq j$ and $G_{(k)} - G$ negative for $k < j$. If a response pattern follows a Guttman pattern there will be more zeros for $k \leq j$ (i.e. harder items) and more ones for $k \geq j$ (i.e. easier items). It is easiest to see looking at (4.34), but this implies that $\xi_1$ would be negative. For the opposite of a Guttman pattern, meaning having more zeros for $k \geq j$ and ones for $k \leq j$ (i.e. incorrect answers to easy items and correct answers to more difficult items), then $\xi_1$ would be positive. A similar argument holds for $\xi_2$ if in (4.39) we consider the descending order of the terms $p_{ik} - T_i$.

It needs to be noted that just because a value of $\xi_1$ or $\xi_2$ is positive does not indicate an aberrant response pattern. Tatsuoka (1984) mention that the exact sampling distribution of $\xi_1$ or $\xi_2$ has not been derived algebraically, but “goodness-of-fit tests of $\xi_1$ and $\xi_2$ with normal distributions provide satisfactory evidence that they may follow approximately normal distributions (Tatsuoka and Tatsuoka 1982).” In practice then, we may use a standard
normal distribution to judge the extremeness of the $\xi_1$ and $\xi_2$ measures.

4.4 Example Using Classical Person Fit Measures

Here we will show an example of how each statistic in section 4.3 is used in a classical setting. We start by generating an $I \times K$ matrix $Y$ of responses of $I = 200$ examinees to $K = 20$ items. The response matrix $Y$ is generated under the 2PN IRT model given in the beginning of the chapter in section 4.1, where the data generating values of the ability parameters $\theta_i$, $i = 1, \ldots, 200$, were generated from a $N(0,1)$ distribution, the difficulty parameters $b_k$, $k = 1, \ldots, 20$, from a $N(0,1)$ distribution, and the discrimination parameters $a_k$, $k = 1, \ldots, 20$, from a $TN_{(0,\infty)}(1,.50)$ distribution. The binary responses $y_{ik}$ were then generated such that $y_{ik} = 1$ with probability $p_{ik} = \Phi(a_k\theta_i - b_k)$ and $y_{ik} = 0$ with probability $1 - p_{ik}$.

To illustrate the statistics’ abilities to detect aberrant response patterns we have simulated the data set $Y$ such that five examinees were guessers. Specifically, examinees 10, 50, 90, 130, and 170. We will discuss later how we will define different types of response violations, i.e. cheating, guessing, etc., in section 4.7.1.1. For now, we assume a guesser to be someone of low ability who tends to get more difficult items correct. In this, the person is assumed to try honestly on easy and moderately difficult items, but guesses the answer to more difficult ones. In simulating this we first simulate the guessing examinees’ abilities from a $U(-2, -.50)$ distribution. Then for items with difficulty parameter greater than one, their responses were generated with a probability of success of 0.70. For all items with difficulty parameter less than one, the examinees’ responses were generated with a probability of success $p_{ik} = \Phi(a_k\theta_i - b_k)$.

In the classical setting, the person fit statistics are evaluated using the maximum likelihood estimates of the parameters in their respective formulas. Using the simulated data set $Y$, we compute the JMLE’s of the item and ability parameters. Using these, we compute the estimated $200 \times 20$ matrix $\hat{P}$ whose entries $\hat{p}_{ik} = \Phi(\hat{a}_k\hat{\theta}_i - \hat{b}_k)$, $i = 1, \ldots, 200$, $k = 1, \ldots, 20$, 

Table 4.3 The first column lists the item number, and the second and third column list the JMLE’s for the item discrimination and difficulty parameters, respectively. The fourth column lists the response pattern for guessing examinee 50 and the last column lists the estimated probabilities $\hat{p}_{50k} = \Phi(\hat{a}_k \theta_{50} - \hat{b}_k)$ (for $\theta_{50} = -0.231$) of success for examinee 50 on each item $k = 1, \ldots, 20$. The values of $\hat{a}_k$ and $\hat{b}_k$, $k = 1, \ldots, 20$, for the discrimination and difficulty parameters, respectively. The fourth column lists the vector of responses of examinee 50 and the last column gives the vector of estimated probabilities $\hat{p}_{50k} = \Phi(\hat{a}_k \hat{\theta}_{50} - \hat{b}_k)$ of success for examinee $i = 50$ on each item $k = 1, \ldots, 20$.

<table>
<thead>
<tr>
<th>item</th>
<th>$\hat{a}$</th>
<th>$\hat{b}$</th>
<th>$Y_{50}$</th>
<th>$\hat{p}_{50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.931</td>
<td>0.629</td>
<td>0</td>
<td>0.199</td>
</tr>
<tr>
<td>2</td>
<td>0.022</td>
<td>0.994</td>
<td>1</td>
<td>0.159</td>
</tr>
<tr>
<td>3</td>
<td>0.523</td>
<td>1.422</td>
<td>1</td>
<td>0.061</td>
</tr>
<tr>
<td>4</td>
<td>1.258</td>
<td>1.191</td>
<td>0</td>
<td>0.069</td>
</tr>
<tr>
<td>5</td>
<td>1.800</td>
<td>0.134</td>
<td>0</td>
<td>0.291</td>
</tr>
<tr>
<td>6</td>
<td>1.028</td>
<td>1.177</td>
<td>0</td>
<td>0.078</td>
</tr>
<tr>
<td>7</td>
<td>1.036</td>
<td>0.663</td>
<td>1</td>
<td>0.183</td>
</tr>
<tr>
<td>8</td>
<td>0.264</td>
<td>-0.497</td>
<td>0</td>
<td>0.669</td>
</tr>
<tr>
<td>9</td>
<td>0.287</td>
<td>-2.011</td>
<td>1</td>
<td>0.974</td>
</tr>
<tr>
<td>10</td>
<td>1.201</td>
<td>-0.139</td>
<td>0</td>
<td>0.445</td>
</tr>
<tr>
<td>11</td>
<td>0.920</td>
<td>0.733</td>
<td>0</td>
<td>0.172</td>
</tr>
<tr>
<td>12</td>
<td>1.159</td>
<td>-0.301</td>
<td>0</td>
<td>0.513</td>
</tr>
<tr>
<td>13</td>
<td>1.405</td>
<td>-0.117</td>
<td>1</td>
<td>0.417</td>
</tr>
<tr>
<td>14</td>
<td>1.625</td>
<td>0.282</td>
<td>1</td>
<td>0.255</td>
</tr>
<tr>
<td>15</td>
<td>0.707</td>
<td>2.085</td>
<td>0</td>
<td>0.012</td>
</tr>
<tr>
<td>16</td>
<td>0.399</td>
<td>1.422</td>
<td>1</td>
<td>0.065</td>
</tr>
<tr>
<td>17</td>
<td>0.762</td>
<td>-1.110</td>
<td>0</td>
<td>0.825</td>
</tr>
<tr>
<td>18</td>
<td>2.197</td>
<td>-0.089</td>
<td>0</td>
<td>0.337</td>
</tr>
<tr>
<td>19</td>
<td>2.819</td>
<td>-0.665</td>
<td>1</td>
<td>0.505</td>
</tr>
<tr>
<td>20</td>
<td>1.382</td>
<td>-0.315</td>
<td>0</td>
<td>0.498</td>
</tr>
</tbody>
</table>
The first statistic we compute for examinee 50 is the $L_z$ statistic. Using the values in the last two columns of table 4.3 we start by computing the values of $L_{0,50}$, $E(L_{0,50})$, and $Var(L_{0,50})$,

\[ L_{0,50} = \sum_{k=1}^{20} [y_{50k}\log(\hat{p}_{50k}) + (1 - y_{50k})\log(1 - \hat{p}_{50k})] \]
\[ = -18.187, \quad (4.40) \]

\[ E(L_{0,50}) = \sum_{k=1}^{20} [\hat{p}_{50k}\log(\hat{p}_{50k}) + (1 - \hat{p}_{50k})\log(1 - \hat{p}_{50k})] \]
\[ = -9.411, \quad (4.41) \]

\[ Var(L_{0,50}) = \sum_{k=1}^{20} \hat{p}_{50k}[1 - \hat{p}_{50k}] \left[ \log \frac{\hat{p}_{50k}}{1 - \hat{p}_{50k}} \right]^2 \]
\[ = 4.646, \quad (4.42) \]

and so the $L_z$ value for examinee 50 is,

\[ L_{z,50} = \frac{-18.187 - (-9.411)}{\sqrt{4.646}} \]
\[ = -4.072. \quad (4.43) \]

In practice, the reference distribution for the $L_z$ statistic is the $N(0, 1)$ distribution. Looking at the calculated value of $L_{z,50} = -4.072$, we see that this is very unlikely if the responses of examinee 50 were generated entirely under the 2PN model. This indicates that the response pattern is outlying under the model.

The first of the residual based statistics is the unweighted mean square fit statistic $U$. 
We compute this for the 50th examinee,

\[ U_{50} = \frac{1}{20} \sum_{k=1}^{20} \frac{(y_{50k} - \hat{p}_{50k})^2}{\hat{v}_{50k}} \]

\[ = 2.795, \quad (4.44) \]

where \( \hat{v}_{50k} = \hat{p}_{50k}(1 - \hat{p}_{50k}) \) is the estimated variances for each of the 20 responses. With this we compute the cube root transformation from (4.13). The standard deviation of \( U_{50} \) is computed from (4.14) as,

\[ S_{U_{50}} = \left[ \sum_{k=1}^{20} \hat{v}_{50k}^{-1} - 4(20) \right] / 20 \]

\[ = 0.669, \quad (4.45) \]

and the transformed value is computed as

\[ t(U_{50}) = \left( U_{50}^{1/3} - 1 \right) \frac{3}{S_{U_{50}}} + \frac{S_{U_{50}}}{3} \]

\[ = \left( (2.795)^{1/3} - 1 \right) \frac{3}{0.669} + \frac{0.669}{3} \]

\[ = 2.056. \quad (4.46) \]

Referencing the transformed value \( t(U_{50}) \) to a \( N(0, 1) \) distribution we see that a value of 2.056 would be extreme.

For the weighted mean square fit statistic \( W \), we compute the value of the measure from (4.15),

\[ W_{50} = \frac{\sum_{k=1}^{20} (y_{50k} - \hat{p}_{50k})^2}{\sum_{k=1}^{20} \hat{v}_{50k}} \]

\[ = 2.057. \quad (4.47) \]
We compute the standard deviation of the measure from (4.16),

\[
S_{W_{50}} = \left[ \sum_{k=1}^{20} \hat{v}_{50k} - 4 \sum_{k=1}^{20} \hat{v}_{50k}^2 \right]^{1/2} \sum_{k=1}^{20} \hat{v}_{50k}
\]

\[= 0.263 \]

We apply the cube-root transformation \(t\) and get,

\[
t(W_{50}) = \left( W_{50}^{1/3} - 1 \right) \frac{3}{S_{W_{50}}} + \frac{S_{W_{50}}}{3}
\]

\[= ((2.057)^{1/3} - 1) \frac{3}{0.263} + \frac{0.263}{3}
\]

\[= 3.188.\]

We see this value of 3.188 is extreme under the standard normal distribution which indicates that examinee 50’s response pattern is unusual under the model.

To calculate the unweighted between fit statistic \(UB\), we need to divide the items into \(S\) non-overlapping subsets. We divide the items into \(S = 2\) subsets of equal size 10, denoted \(A_1\) and \(A_2\). We do this by first ranking the items based on their difficulty estimates \(\hat{b}_k\), \(k = 1, \ldots, 20\), and then splitting the ranked items in half. The first half of the items with the lowest difficulty estimates will make up set \(A_1\), and set \(A_2\) will be comprised of the second half.

Table 4.4 gives the items listed in rank order according to the item difficulty mle’s \(\hat{b}_k\), \(k = 1, \ldots, 20\). In the top two rows, the first row indicates the items that are in the first subset \(A_1\), and the second row gives their corresponding ranked estimates \(\hat{b}_{(k)}\). In the bottom two rows, the top row gives the items that are in the second subset \(A_2\), and the second row gives their corresponding ranked difficulty parameter estimates.

We calculate the numerator and denominator in (4.27) using values in the last two columns in table 4.3 partitioned according to the items in sets \(A_1\) and \(A_2\). For the items in
Table 4.4 The items are listed in rank order according to their corresponding mle $\hat{b}_k$, $k = 1, \ldots, 20$. The top 2 rows give the first set of items $A_1$ that compose the bottom half of the items, and the ranked difficulty estimates $\hat{b}_{(k)}$. The bottom 2 rows give the second set of items $A_2$ making up the top half and their difficulty estimates.

<table>
<thead>
<tr>
<th>$A_1$</th>
<th>9</th>
<th>17</th>
<th>19</th>
<th>8</th>
<th>20</th>
<th>12</th>
<th>10</th>
<th>13</th>
<th>18</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_{(k)}$</td>
<td>-2.011</td>
<td>-1.11</td>
<td>-0.665</td>
<td>-0.497</td>
<td>-0.315</td>
<td>-0.301</td>
<td>-0.139</td>
<td>-0.117</td>
<td>-0.889</td>
<td>0.134</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$A_2$</th>
<th>14</th>
<th>1</th>
<th>7</th>
<th>11</th>
<th>2</th>
<th>6</th>
<th>4</th>
<th>16</th>
<th>2</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_{(k)}$</td>
<td>0.282</td>
<td>0.629</td>
<td>0.663</td>
<td>0.733</td>
<td>0.994</td>
<td>1.177</td>
<td>1.191</td>
<td>1.422</td>
<td>1.422</td>
<td>2.085</td>
</tr>
</tbody>
</table>

For the items in set $A_1$ we compute,

$$\left[ \sum_{j \in A_1} (y_{50j} - \hat{p}_{50j}) \right]^2 = 6.113,$$

(4.50)

and

$$\sum_{j \in A_1} \hat{p}_{50j}(1 - \hat{p}_{50j}) = 2.061.$$  

(4.51)

For the items in set $A_2$ we compute,

$$\left[ \sum_{j \in A_2} (y_{50j} - \hat{p}_{50j}) \right]^2 = 14.042,$$

(4.52)

and

$$\sum_{j \in A_1} \hat{p}_{50j}(1 - \hat{p}_{50j}) = 1.042.$$  

(4.53)

Then the $UB$ statistic is then computed as,

$$UB_{50} = \frac{1}{2 - 1} \left[ \frac{6.113}{2.061} + \frac{14.029}{1.042} \right]$$

$$= 16.430.$$  

(4.54)

In [Smith (1985)] a critical value of 2 was used for judging the extremeness of the $UB$ statistic. We see that our calculated value of $UB_{50} = 16.430$ is much larger.

To compute the statistics $\xi_1$ and $\xi_2$ we find the column means $G_k$, $k = 1, \ldots, 20$, and row mean $T_{50}$ of the 50th row of the estimated probability matrix $\hat{P}$. The values of $G_k$,
\(k = 1, \ldots, 20\), are given in Table 4.5. The row mean \(T_{50} = 0.336\) can be found by taking the mean of the last column of values in Table 4.3, and the overall mean \(G = 0.421\) is found by taking the mean of the \(G_k, k = 1, \ldots, 20\).

<table>
<thead>
<tr>
<th>item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>(G_k)</td>
<td>0.323</td>
<td>0.160</td>
<td>0.104</td>
<td>0.233</td>
<td>0.467</td>
<td>0.209</td>
<td>0.323</td>
<td>0.684</td>
<td>0.973</td>
<td>0.529</td>
</tr>
<tr>
<td>item</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>(G_k)</td>
<td>0.295</td>
<td>0.571</td>
<td>0.520</td>
<td>0.436</td>
<td>0.045</td>
<td>0.093</td>
<td>0.810</td>
<td>0.504</td>
<td>0.573</td>
<td>0.565</td>
</tr>
</tbody>
</table>

Table 4.5 Column means \(G_k, k = 1, \ldots, 20\), of the estimated probability matrix \(\hat{P}\).

To find the value of the \(\xi_1\) statistic we compute the numerator and denominator of (4.35). That is,

\[
\sum_{k=1}^{20} (G_k - G)(\hat{p}_{50k} - y_{50k}) = 1.404, \\
(4.55)
\]

and

\[
\left(\sum_{k=1}^{20} \hat{\sigma}_{50k}^2 (G_k - G)^2\right)^{1/2} = 0.320, \\
(4.56)
\]

where \(\hat{\sigma}_{50k}^2 = \hat{p}_{50k}(1 - \hat{p}_{50k})\) with the values \(\hat{p}_{50k}, k = 1, \ldots, 20\), being in the last column of Table 4.3. We then compute the value of \(\xi_1\) as,

\[
\xi_{1,50} = \frac{1.404}{0.320} = 4.389. \\
(4.57)
\]

Likewise, for \(\xi_2\) we compute the numerator and denominator in (4.37),

\[
\sum_{k=1}^{20} (\hat{p}_{50k} - T_{50})(\hat{p}_{50k} - y_{50k}) = 1.432, \\
(4.58)
\]

and

\[
\left(\sum_{k=1}^{20} \hat{\sigma}_{50k}^2 (\hat{p}_{50k} - T_{50})^2\right)^{1/2} = 0.362. \\
(4.59)
\]
The value of $\xi_2$ for examinee 50 is then computed as,

$$
\xi_{2,50} = \frac{1.432}{0.362} = 3.956,
$$

(4.60)

If we use the $N(0, 1)$ distribution as a reference then we see that both the values of 4.389 and 3.956 are extreme.

### 4.4.1 Discussion on Person Fit Measures

We have already mentioned a little bit on how some of these measures are useful in detecting aberrant response patterns. The measures $\xi_1$ and $\xi_2$ should be effective in detecting when an examinee answers easy items incorrectly and more difficult items correctly. The statistics $U$, $W$, and $UB$ directly compute the discrepancy between the observed item scores and expected item scores so these should detect patterns where responses are not in accordance with the model. Likewise, $L_z$ will do the same, but afford greater interpretability of the statistic value to classify a response as aberrant or not. Overall, the $L_z$ statistic is one of the most popular person fit statistics due to its usefulness and also because it has been shown to be very powerful in detecting misfitting item response patterns [Nering and Meijer 1998, Drasgow et al. 1987].

Out of so many person fit statistics, it would be nice to identify which one works the best. The problem is that the types of response violations (i.e. guessing, cheating, etc.) can vary. What has become evident from the research that has been done is that the statistics may perform better at detecting certain types of response violations than others. This has been investigated by some researchers in which the detection rates of measures for certain types of aberrant behaviors were compared. Karabatsos (2003) performed a simulation study comparing, amongst others, the six measures we have discussed. Their power was investigated under the Rasch model in detecting *cheating, careless responding*, where an
examinee mistakenly answers easy items incorrectly, *random responding*, where items are answered in such a way that the probability of success is constant, and *creative responding*, where students may answer easy items incorrectly due to misinterpreting the information in the question. As can be seen in the study, the measures performed differently for different types of aberrant behavior.

For example, of the six measures we have discussed the $U$ statistic had one of the highest rates for detecting random responding and cheating. However, it had a relatively low rate for detecting creative responding. Likewise, the $L_z$ statistic did relatively well in detecting guessing, but not as well in detecting say cheating or creative responding. The point to take from this is that there is not one “best” person fit statistic. What we can realistically try to do is identify statistics that perform reasonably well over a multitude of response violation types.

### 4.5 Checking Person Fit Using Bayesian Model Checking Methods

An alternative to classical IRT person fit analysis can be done from a Bayesian perspective. In this section we will discuss how the three Bayesian model checking methods: posterior predictive (PP), prior predictive posterior simulation (PPPS), and pivotal discrepancy measure (PDM) can be used to determine person fit in a Bayesian 2PN IRT model. We start with the PP method and then discuss the PPPS and PDM methods.

We will be using the person fit measures discussed in sections 4.3 and 4.4 under the three methods. How these will be used under each Bayesian method will be described when talking about each method. Before we do this however, we will introduce another measure that we will use under these three Bayesian methods. This measure is based on the latent variables $Z_{ik}$ introduced in chapter 2.
4.5.1 Bayesian Latent Residual as a Person Fit Measure

Here we discuss how the latent response variables $Z_{ik}$ can be used to compute a person fit discrepancy measure. Recall that Albert (1992) introduced augmented latent variables into the 2PN IRT model, where $Z_{ik} \sim N(a_k \theta_i - b_k, 1)$. Albert and Chib (1995) introduced using residuals based on the latent variables, called Bayesian latent residuals, for outlier detection in binary response regression models. Here, an outlier is an observed score $y$ which is in conflict with the probability of success under the model. These were introduced as an alternative to Bayesian residuals and a thorough discussion can also be found in Johnson and Albert (1999).

We will give a brief discussion on Bayesian residuals in IRT models and how they are used for outlier detection. Likewise, we will then describe what Bayesian latent residuals are and how they are used. Then we will discuss how Bayesian latent residuals can be used as a person fit measure for IRT under the three model checking methods.

4.5.1.1 Bayesian Residuals

Using a 2PN IRT model as an example, a Bayesian residual is defined as,

$$r_{ik} = y_{ik} - \Phi(a_k \theta_i - b_k).$$

(4.61)

The marginal posterior density of the residual, $\pi(r_{ik}|y)$, is continuous-valued and the posterior mean is used to estimate the residual value. The marginal posterior of $r_{ik}$ is defined on different domains depending on the value of $y_{ik}$. If $y_{ik} = 1$ then the marginal posterior of $r_{ik}$ is defined on the interval $(0,1)$, and on $(-1,0)$ if $y_{ik} = 0$. Values of the residuals that are far from zero (close to $-1$ or 1) are taken to indicate outlying observations. The marginal posteriors of the Bayesian residuals also depend on the distributions of the parameters in the model, but are typically unknown analytically. This, along with the fact that they can have different domains can make it difficult to assess extremeness of the densities. The prior
distribution of $r_{ik}$ is not known either, due to the discrete nature of $y$, and this makes it hard to understand extreme sizes of the residual a priori. All this in turn makes it difficult to assess outliers.

### 4.5.1.2 Bayesian Latent Residuals

Introduced in chapter 1, the latent data $Z_{ik}$ is viewed as a continuous latent variable that determines the performance of examinee $i$ on item $k$. The item is believed to be answered correctly if $Z_{ik} > 0$, and incorrectly if $Z_{ik} \leq 0$. The observed response data is thought of as indicator variables for $Z_{ik}$, that is $y_{ik} = 1$ if $Z_{ik} > 0$ and $y_{ik} = 0$ otherwise. For explanation we still assume the setting of a 2PN IRT model, incorporating a latent variable $Z$. In this model the prior distribution of the latent variable $Z_{ik}$ is $N(m_{ik},1)$, where $m_{ik} = a_k\theta_i - b_k$. The conditional posterior distribution of $Z_{ik}$ is truncated normal with mean $m_{ik}$ and variance 1. In particular, the conditional posterior of $Z_{ik}$ is $TN_{(-\infty,0)}(m_{ik},1)$ if $y_{ik} = 0$ and $TN_{(0,\infty)}(m_{ik},1)$ if $y_{ik} = 1$.

As mentioned, Bayesian latent residuals were introduced in [Albert and Chib (1995)](1995) for binary response regression models. [Fox (2010)](2010) and [Toribio (2006)](2006) discuss these in an IRT model setting. A Bayesian latent residual is the difference between the latent variable and the mean $m_{ik}$. Here, we define the Bayesian latent residual as,

$$
\epsilon_{ik} = Z_{ik} - m_{ik}, \quad (4.62)
$$

for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. Based on this we can see that the latent residual $\epsilon_{ik}$ has a standard normal prior distribution. We also know the conditional posterior distribution of $\epsilon_{ik}$ to be truncated standard normal, specifically $TN_{(-\infty,-m_{ik})}(0,1)$ if $y_{ik} = 0$ and $TN_{(-m_{ik},\infty)}(0,1)$ if $y_{ik} = 1$.

In these posteriors we see that the point of truncation is $-m_{ik}$. If the observed data are not in conflict with the fitted probabilities of success $\Phi(m_{ik})$, then the latent residual posterior
distributions are close to standard normal curves. That is, if \( y_{ik} = 1 \) with fitted probability \( \Phi(m_{ik}) \) close to 1, or \( y_{ik} = 0 \) with \( \Phi(m_{ik}) \) close to 0. When the fitted probability of success is in conflict with the observed response then the posterior distribution is significantly different from a standard normal distribution. We see this if we look close at the truncation point. For \( y_{ik} = 0 \) the truncation is done to the left of \(-m_{ik}\). If \( m_{ik} \) is positive (implying \( \Phi(m_{ik}) \) closer to 1) the posterior distribution does not include zero. If \( m_{ik} \) is negative (implying \( \Phi(m_{ik}) \) closer to 0) then zero is in the posterior of \( \epsilon_{ik} \). Similarly, if \( y_{ik} = 1 \) the truncation is done to the right of \(-m_{ik}\) and if \( m_{ik} \) is negative then zero is not in the posterior of \( \epsilon_{ik} \). Zero is in the posterior if \( m_{ik} \) is positive when \( y_{ik} = 1 \). Therefore, the posterior distribution of \( \epsilon_{ik} \) will not include zero when \( y_{ik} = 0 \) and \( m_{ik} > 0 \) or \( y_{ik} = 1 \) and \( m_{ik} < 0 \). Hence, latent residuals far from zero are considered outlying and imply outlying responses under the model.

4.5.1.3 Illustration of Bayesian Residuals and Bayesian Latent Residuals

As an example, under the 2PN IRT model from section 4.1 we simulate a data matrix \( Y \) of binary responses to an exam with \( I = 200 \) examinees and \( K = 20 \) items. To illustrate use of these residuals in detecting outlying responses, the data were simulated such that some of the examinees were guessers. Guessing was simulated by taking examinees with low ability and having a high probability of correct response to difficult items. Specifically, the ability of guessing examinees was simulated from a \( U(-2, -0.5) \) distribution and for every item with difficulty parameter greater than 1 (representing a difficult item) the probability of success was set to 0.70.

We fit the model using MCMC and use the output to compute the Bayesian residuals for the item responses for two examinees, a guesser and a non-guesser. We generated density plots of the marginal posterior distributions of the residuals and also 90% probability interval plots. These are given in figure 4.1. The density plots are plotted over the intervals \((-1, 0)\) or \((0, 1)\). The 90% probability interval plots are plotted as box plots with center line signifying
the posterior median and outer fences extending to the 5th and 95th quantiles. The box plots are plotted from left to right by increasing difficulty of the items. In this, the posterior means of the difficulty parameters were used as estimates and the items are ordered accordingly. The plots on the top are of a non-guessing examinee. We see that none of the distributions are extremely close to $-1$ or $1$, indicating that none of the responses are in obvious conflict with the probability of success under the model. In fact, we see that the response pattern follows a Guttman pattern as is expected under the model. We see this looking at the 90% probability interval plot that the answers to easier items are correct (positive residuals) and as items get more difficult the answers are more incorrect (negative residuals).

The plots on the bottom are for a guessing examinee. From the 90% probability interval plot we see that this corresponds to three of the most difficult items which were correctly answered while the rest of the responses follow under the model (i.e. correct for easier items and incorrect for more difficult items). The ability estimate for this examinee was $\theta = -0.76$ which is relatively low, so correct responses to the three difficult items were unusual under the model. We see that the densities of these residuals are very close to one, indicating these responses are in conflict with the model. However, the “How close is close?” question remains when interpreting all of these. For example, look at the box plot in the top right plot for item 7 for examinee 3 and the box plot in the bottom right for item 20 for examinee 10. These show the posteriors are away from zero, but not extremely close to one either. So do these signify outliers or not? An easier way of classifying residuals as outlying is what Bayesian latent residuals give us.

Using the simulated data and MCMC output in section 4.5.1.1 we compute the Bayesian latent residuals corresponding to the item responses for the same two examinees. We plot the posterior densities and 90% probability intervals for the posterior distributions of the latent residuals in figure 4.2. The density plots are shown with a standard normal curve superimposed for reference. The 90% probability intervals are again plotted as box plots from the 5th to the 95th quantiles of the posterior samples.
Figure 4.1 Density and 90% probability interval plots for the Bayesian residuals corresponding to the responses of the 20 items by a non-guessing examinee (top) and a guessing examinee (bottom). Posterior distributions very close to −1 or 1 indicate outlying responses.
Figure 4.2 Density and 90% probability interval plots for the Bayesian latent residuals corresponding to the responses of the 20 items by a non-guessing examinee (top) and a guessing examinee (bottom). A standard normal density curve is imposed in the density plots for reference. Posterior distributions significantly far from zero correspond to outlying responses.
The posterior means of the latent residuals are negative when $y_{ik} = 0$ and positive if $y_{ik} = 1$. Thus, the location of the center of the distribution, above zero or below zero, indicates if the observed response was correct or incorrect, respectively. How extreme this shift above or below zero is indicates where the observed response is in agreement with the model. The box plots help indicate this a little easier. For these, if zero is outside the box plot then the residual is considered outlying.

The plots on top are for the non-guessing examinee. We see that most of the posterior densities in the top left are close to standard normal curves, except for two which are shifted away from zero. Similarly, we see in the top right that most of the box plots contain zero. Again, if the median is above (below) zero this indicates the observed response was correct (incorrect). We see that, for example, the box plot in the top right plot corresponding to item 7 does not contain zero and is negative. This shows that the examinee’s response to item 7 was incorrect, but was unexpected given their ability level and how easy the item was.

The plots on the bottom are for the guessing examinee. In the density plots we definitely see three densities shifted far to the right of zero. These correspond to the three difficult items that were answered correctly. Likewise, we see the box plots in the bottom left for the latent residuals of these three items which are positive and do not include zero. Both of these indicate that the responses were aberrant under the model.

Going back to the question at the end of section 4.5.1.1 we can see how latent residuals are an improvement over Bayesian residuals. That is, for the box plot of item 7 for examinee 3 and item 20 for examinee 10 in figure 4.1 it was not clear if these would indicate outlying responses. Looking at the corresponding box plots in figure 4.2 we see that both box plots do not include zero. Thus, using latent residuals we have clear evidence that these residuals are extreme and the corresponding observed responses are outlying.
4.5.1.4 Using Bayesian Latent Residuals as a Person Fit Measure

As was just discussed, the latent residuals can be used to assess which responses to items are aberrant under the assumed model. Person fit deals with how aberrant an examinee’s item responses are as a whole under the model. Using the latent residuals corresponding to the item responses for an examinee we can create a summary measure over all items by summing the squared latent residuals for each item. That is, we can define the measure,

\[ d_{Z_i} = \sum_{k=1}^{K} \epsilon_{ik}^2. \] (4.63)

Now, since the prior distribution of the latent residuals is \( N(0, 1) \) then the prior distribution the measure \( d_{Z_i} \) is \( \chi^2_K \), for \( i = 1, \ldots, I \). This gives some idea \textit{a priori} what the scale of this measure is. The posterior distribution is not known because the exact distribution of the sum of squares of truncated normal random variables is not known. However, in section 4.5.1.2 we saw that the less the conflict between the observed responses and the model, the closer the latent residual posterior distributions are to standard normal. This implies the posterior of \( d_{Z_i} \) should be close to a \( \chi^2_K \) distribution.

Likewise, outlying responses have corresponding latent residuals that are significantly different from zero. This implies the sum of squared residuals over a response pattern with outlying responses would be large. In particular then, values from the posterior distribution of \( d_{Z_i} \) should be greater than from a \( \chi^2_K \) distribution. To illustrate this we compute a posterior sample of the measure \( d_{Z_i} \) for examinee \( i = 3 \) (non-guesser) and examinee \( i = 10 \) (guesser) from above using the simulated latent residuals from section 4.5.1.3 and plot a histogram of the posterior sample with a \( \chi^2_K \) density curve superimposed for reference.

Specifically, a posterior sample for \( d_{Z_i} \) can be computed using the posterior samples of latent residuals computed in section 4.5.1.3 \( (\tilde{\epsilon}_{ik}^{(1)}, \ldots, \tilde{\epsilon}_{ik}^{(L)}) \), for \( k = 1, \ldots, K \). Then a posterior sample of \( d_{Z_i} \) of length \( L \) can be computed by summing the squared posterior latent residuals over \( k \), for \( l = 1, \ldots, L \). That is, we compute \( \tilde{d}_{Z_i}^{(l)} = \sum_{k=1}^{K} \tilde{\epsilon}_{ik}^{(l)} \), for \( l = 1, \ldots, L \). We
give the plots of the computed posterior samples of $d_{Z_i}$ for examinee $i = 3$ and $10$ in figure 4.3.

Figure 4.3 Posterior samples of person fit measure $d_Z$ for examinee 3 (non-guesser) and examinee 10 (guesser) are graphed in the histograms. A $\chi^2_K$ density curve is superimposed for reference. We see that the posterior sample for examinee 3, whose responses were generated under the model is closer to a $\chi^2_K$ distribution than that of examinee 10 whose responses were simulated with guessing involved. The posterior sample of $d_Z$ for examinee 10 is notably shifted to the right.

We see in the plots of figure 4.3 that the posterior sample of $d_Z$ for examinee 3 is closer to a $\chi^2_K$ distribution than that of examinee 10 which is shifted farther to the right. This is since the item responses of examinee 3 were generated completely under the model and the item responses of examinee 10 were generated with guessing involved. As we saw in the bottom right of figure 4.2, the aberrant guessed responses had unusually large latent residuals. This contributes to the sum of squared latent residuals $d_Z$ being larger.

So we see that the measure $d_Z$ can indicate if an examinee has given aberrant responses to items. We can use this measure to assess the fit of an examinee’s response pattern, but what is needed is a reference distribution for which to compare the computed posterior values of $d_Z$ to judge extremeness. We can graphically compare the posterior sample of $d_Z$ to a $\chi^2_K$ distribution as in figure 4.3 as an informal check. However, it is more useful to have a more formal way to judge if the values of $d_Z$ are extreme. This can be done using the three
model checking methods discussed already: PP, PDM and PPPS. We will discuss how the $d_Z$ measure can be used under each of these methods in the following sections.

We will now return to discuss how the PP, PPPS, and PDM model checking methods can be used in an IRT setting. The person fit measures we will be using are the six classical measures previously discussed and the $d_Z$ measures just introduced. We will give a discussion of how the methods can be used and some of the advantages they offer over the classical approach. An in-depth example showing how each person fit measure is computed under each method will be given later in section 4.6.

4.5.2 Posterior Predictive Method

As mentioned, the posterior distribution of the parameters of a Bayesian IRT model can be simulated from using a Gibbs MCMC sampling algorithm as outlined in section 2.5.2. So, for a simulated sample from the posterior distribution of ability and item parameters, $(\tilde{\theta}^{(1)}, \tilde{\xi}^{(1)}), \ldots, (\tilde{\theta}^{(N)}, \tilde{\xi}^{(N)})$ where,

$$\tilde{\theta}^{(n)} = \left( \tilde{\theta}_1^{(n)}, \ldots, \tilde{\theta}_I^{(n)} \right)$$ (4.64)

and

$$\tilde{\xi}^{(n)} = \left( \tilde{\alpha}_1^{(n)}, \tilde{\beta}_1^{(n)}, \ldots, \tilde{\alpha}_K^{(n)}, \tilde{\beta}_K^{(n)} \right),$$ (4.65)

we can simulate posterior predictive response data $\tilde{Y}_{rep}^{(n)}$ from the posterior predictive distribution $f(\tilde{Y}_{rep} | \tilde{\theta}^{(n)}, \tilde{\xi}^{(n)})$, for $n = 1, \ldots, N$. We do this for each $n = 1, \ldots, N$ by simulating each entry $\tilde{y}_{ik}^{(n)}$ as a Bernoulli random variable with probability of success $\tilde{p}_{ik}^{(n)} = \Phi(\tilde{\alpha}_k^{(n)} \tilde{\theta}_i^{(n)} - \tilde{\beta}_k^{(n)})$ for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. From the general description in chapter 2, we can effectively compute a sample from the posterior predictive distribution of a given person fit measure $T$ using the posterior predictive data sets $\tilde{Y}_{rep}^{(n)}$. Since $T$ is a person fit measure we calculate a posterior predictive sample for each examinee $i = 1, \ldots, I$. 
say,

\[ T_i \left( \tilde{y}_i^{(1)} \left( \tilde{\theta}_i^{(1)}, \tilde{\xi}^{(1)} \right) \right), \ldots, T_i \left( \tilde{y}_i^{(N)} \left( \tilde{\theta}_i^{(N)}, \tilde{\xi}^{(N)} \right) \right), \]  

(4.66)

where \( \tilde{y}_i^{(n)} \) is the response vector of the \( i^{th} \) examinee from \( \tilde{Y}_{rep} \).

Then the measure is evaluated using the observed data and posterior sample of parameters,

\[ T_i \left( y_{i,obs} \left( \tilde{\theta}_i^{(1)}, \tilde{\xi}^{(1)} \right) \right), \ldots, T_i \left( y_{i,obs} \left( \tilde{\theta}_i^{(N)}, \tilde{\xi}^{(N)} \right) \right), \]  

(4.67)

for \( i = 1, \ldots, I \), where \( y_{i,obs} \) is the observed response vector of the \( i^{th} \) examinee. With these, the relevant pair-wise proportion of posterior predictive sample values that are greater than those computed using the observed data is found. That is, for each \( i = 1, \ldots, I \) the proportion

\[
\frac{\sum_{n=1}^{N} I \left[ T_i \left( \tilde{y}_i^{(n)} \left( \tilde{\theta}_i^{(n)}, \tilde{\xi}^{(n)} \right) \right) \geq T_i \left( y_{i,obs} \left( \tilde{\theta}_i^{(n)}, \tilde{\xi}^{(n)} \right) \right) \right]}{N}
\]  

(4.68)

where \( I[\cdot] \) is an indicator function. This is used to estimate the tail area probability

\[
P \left( T_i \left( \tilde{y}_i, \left( \tilde{\theta}_i, \tilde{\xi} \right) \right) \geq T_i \left( y_{i,obs} \left( \tilde{\theta}_i, \tilde{\xi} \right) \right) \mid Y_{obs} \right),
\]  

(4.69)

for \( i = 1, \ldots, I \).

The choices of \( T \) that we will use are the person fit measures discussed in section 4.3, namely: \( L_z, U, W, UB, \xi_1, \) and \( \xi_2 \). We will also employ the \( d_Z \) measure, but this will be discussed shortly. Recall that posterior predictive \( p \)-values of a measure near 0 or 1 are taken as evidence of of lack-of-fit of the observed data with the model. For the person fit measures we will take \( p \)-values less than 0.05, for all measures except \( L_z \), to indicate a lack of fit of the \( i^{th} \) examinee’s response vector to the model. When the \( L_z \) measure is computed under the PP method using response data that is aberrant under the model, the \( p \)-values tend to be larger (i.e. closer to 1) rather than smaller (i.e. closer to 0). For this reason, we choose a threshold value on 0.95 when evaluating posterior predictive \( p \)-values of \( L_z \).
If we think about this last part, this makes sense. If the \( p \)-values for \( L_z \) are closer to 1, this means that the observed values of \( L_z \) computed in the above proportion (4.68) are smaller than the ones given the posterior predictive data. Recall that \( L_z \) is a function of the log-likelihood function \( L_0 \). So, having smaller values given the observed data relative to the posterior predictive values implies a smaller likelihood of the observed data occurring under the assumed model. This is exactly what we would expect if the observed data is aberrant under the model.

The advantage of the PP method is that the posterior predictive distribution serves as a reference distribution to compare the measure computed using the observed data and posterior parameter values. This alleviates the issue mentioned in the classical setting of not knowing the sampling distribution of a statistic. For example, the \( U \), \( W \), and \( UB \) statistics were described as “chi-squared like” in that they did not have a true chi-squared distribution when evaluated with data under the model. The transformation of (4.13) was used and the resulting reference distribution was \( N(0, 1) \), but this was even brought into question because estimates of the true values of the parameters had to be used in the calculation.

To illustrate this we look at the \( UB \) statistic. In section 4.3.2 we discussed how the transformation \( t(UB) \) of the \( UB \) statistic of (4.27) is used in a classical setting. The exact sampling distribution of \( t(UB) \) is unknown, but Smith (1985) showed that sample values of \( t(UB) \) were close to standard normal using simulated data under the Rasch model. Under a 1PN IRT model we simulated 500 \( I \times K \) response data sets with \( I = 200 \) examinees and \( K = 40 \) items. For this model we used the 2PN IRT model given in section 4.1 with the restriction that \( a_k \equiv 1 \). For each data set, we computed the JMLE’s of the item and ability parameters and used these to compute values of \( t(UB_i), i = 1, \ldots, I \), just as was was done for our example in section 4.4. For each \( i \) then, we have a sample of 500 values of \( t(UB_i) \). From Smith (1985), these samples should be close to a \( N(0, 1) \) distribution. For three randomly selected examinees, \( i = 39, 85, \) and 147, we plot a histogram of the sample of \( t(UB_i) \) with a \( N(0, 1) \) density curve superimposed for reference. These are given in figure 4.4, where the
samples of $t(UB_i)$ are plotted for $i = 39$, 85, and 147 respectively from left to right.

Figure 4.4 Histograms of 500 sample values of the measure $t(UB_i)$, for three randomly selected examinees $i = 39$, 85, and 147, simulated under the 1PN IRT model. We want to see how close the sampling distribution of the measure is to a $N(0, 1)$ distribution, so a $N(0, 1)$ density curve superimposed for reference.

We see that the three samples of $t(UB_i)$ are close to a $N(0, 1)$ distribution. However, there is some right skewness shown in the plots. This could lead us to believe that statistic $t(UB_i)$ does not have a true $N(0, 1)$ sampling distribution when the data fit the 1PN IRT model.

As mentioned, this problem is taken care of under the PP method. In this, using MCMC simulation we can effectively simulate posterior parameter values and use these to simulate a sample of posterior predictive data. Using both of these we can compute a posterior predictive sample of the measure $UB_i$, $i = 1, \ldots, I$, as in (4.66). Although the posterior predictive distribution of $UB_i$ may be intractable, we know we have a sample from it. We can now evaluate $UB_i$ using the simulated posterior values of the parameters and the observed data $Y_{obs}$ as in (4.67) and use this to estimate the tail area probability in (4.69).

Following along with this, the posterior predictive distribution of a measure can be sampled from for any model for which we can obtain posterior samples of the parameters. This means that statistics whose use was limited to checking person fit in only certain models under the classical approach can now be used for others. For example, the $UB$ statistic whose sampling distribution is not known under the 2PN and 3PN IRT models can be used
to check person fit under these models in a Bayesian approach.

For example, similar to our example above we simulated 500 response data sets of $I = 200$ examinees and $K = 40$ responses under a 2PN IRT model. For each data set, we computed the JMLE’s of the item and ability parameters and used these to compute values of $t(UB_i)$, $i = 1, \ldots, I$. For the same examinees $i = 39$, $85$, and $147$, we we plot a histogram of the sample of $t(UB_i)$ with a $N(0,1)$ density curve superimposed for reference. These are given in figure 4.5, where the samples of $t(UB_i)$ are plotted for $i = 39$, $85$, and $147$ respectively from left to right.

![Sample of t(UB) Under Classical 2PN IRT Model](image)

Figure 4.5 90% probability interval plots of Bayesian latent residuals for 6 examinees identified with miss-fitting response patterns by the $L_x$ person fit measure evaluated using the PP, PPS, and PDM methods.

As in the 1PN case, we do not know the sampling distribution of $t(UB_i)$ when the data follows the 2PN IRT model. We check to see if it might be close to a $N(0,1)$ distribution, but our samples suggest otherwise. We see that each sample is centered slightly to the right of zero and is still skewed slightly to the right. In this classical setting, we would now have the question of what reference distribution we could use. In the PP method, this is not an issue since we can follow the same steps for simulating a posterior predictive sample of the $UB_i$ measure in the 2PN model case as mentioned above.

Posterior predictive checking is very widely used for person fit checking in Bayesian IRT models. However, there are two alternative methods: the prior predictive posterior simulation (PPPS) based approach and the method using pivotal discrepancy measures (PDMs).
Both of these methods, like the posterior predictive checking approach, solve the above mentioned issues.

4.5.2.1 Using the $d_Z$ Measure under the PP Method

Here we briefly explain how the $d_Z$ measure can be used under the PP method. For each examinee $i$, we can compute a posterior predictive sample of the measure $d_Z i$ and use this along with a posterior sample of $d_Z i$ based on the observed data to obtain a posterior predictive $p$-value. We start by obtaining a posterior predictive sample of the Bayesian latent data $Z$. Fox (2010, ch.5) describes a forward simulation method for generating posterior predictive latent data.

In this procedure, we start by simulating posterior draws of the parameters $\tilde{\theta}^{(n)}$ and $\tilde{\xi}^{(n)}$, and latent variables $\tilde{Z}^{(n)}$, $n = 1, \ldots, N$, from their marginal posterior distributions. We note that this is standard output from our MCMC algorithm. Then we replicate data $\tilde{Y}_{rep}^{(n)}$ from posterior predictive density $f(\tilde{Y}_{rep}^{(n)}|\tilde{\theta}^{(n)}, \tilde{\xi}^{(n)})$. Finally, we can draw latent data replicate $\tilde{Z}_{rep}^{(n)}$ from the conditional density $f(Z_{rep}^{(n)}|\tilde{Y}_{rep}^{(n)}, \tilde{\theta}^{(n)}, \tilde{\xi}^{(n)})$, for $n = 1, \ldots, N$. This is done by simulating $Z_{rep,ik}^{(n)} \sim T N(0, \infty) (a_k^{(n)} \tilde{\theta}_i^{(n)} - b_k^{(n)}, 1)$, if $\tilde{y}_{rep,ik}^{(n)} = 1$ and $Z_{rep,ik}^{(n)} \sim T N(-\infty, 0) (a_k^{(n)} \tilde{\theta}_i^{(n)} - b_k^{(n)}, 1)$, if $\tilde{y}_{rep,ik}^{(n)} = 0$.

Then, posterior predictive latent residuals can be computed by $\tilde{\epsilon}_{rep,ik}^{(n)} = \tilde{Z}_{rep,ik}^{(n)} - \tilde{m}_{ik}^{(n)}$ where $\tilde{m}_{ik}^{(n)} = a_k^{(n)} \tilde{\theta}_i^{(n)} - b_k^{(n)}$. We can then compute a posterior predictive sample of the measure $d_{Z_{rep,n}}^{(rep,n)} = \sum_{k=1}^{K} (\tilde{\epsilon}_{rep,ik}^{(n)})^2$. If we also have a posterior sample of the measure $d_{Z_i}^{(n)} = \sum_{k=1}^{K} (\tilde{\epsilon}_{ik}^{(n)})^2 = \sum_{k=1}^{K} (Z_{ik}^{(n)} - \tilde{m}_{ik}^{(n)})^2$ computed from the posterior distribution based on the observed data $Y_{obs}$, then we can estimate the $p$-value,

$$P \left( d_{Z_{rep}}^{(rep)} \geq d_{Z_i}^{(n)} | Y_{obs} \right), \quad (4.70)$$
by computing the pair-wise proportion
\[
\sum_{n=1}^{N} I\left(d_{Z_i}^{(rep,n)} \geq d_{Z_i}^{(n)}\right),
\]
(4.71)
where \(I(.)\) is an indicator function. If the proportion in (4.71) is smaller than some threshold value of say 0.05, this indicates that the response pattern of examinee \(i\) does not fit under the assumed model.

### 4.5.3 Prior Predictive Posterior Simulation (PPPS) Method

We will discuss how the PPPS method can be used for model checking in IRT models. In an IRT setting, our person fit measures are computed for each examinee so the general PPPS procedure that was outlined in chapter 2 will be done for each \(i = 1, \ldots, I\). The end purpose is to find evidence whether a given examinee’s observed response pattern \(y_{i,obs}\) is aberrant under the model.

Given observed data \(Y_{obs}\) and person fit measure \(T\), we draw a sample of the ability and item parameters from the posterior distribution as in (4.64) and (4.65). For each examinee \(i = 1, \ldots, I\), we use this to compute a posterior sample of \(T\) as in (4.67), denoted \(\{T_i^{(0)}\}\).

We also simulate \(M\) replicate data sets, \(Y_{rep}^{(1)}, \ldots, Y_{rep}^{(M)}\), under the prior predictive distribution. Specifically, we simulate each \(Y_{rep}^{(m)}\) by: (1) simulating ability parameters \(\theta_1, \ldots, \theta_I\) from their \(N(0, 1)\) prior distribution, item discrimination parameters \(a_1, \ldots, a_K\) from their \(TN(0, \infty)(1, .50)\) prior distribution, and item difficulty parameters \(b_1, \ldots, b_K\) from their \(N(0, 1)\) prior distribution and (2) simulating dichotomous responses \(y_{ik,rep}^{(m)}\), where \(y_{ik,rep}^{(m)} = 1\) with probability \(\Phi(a_k\theta_i - b_k)\) and \(y_{ik,rep}^{(m)} = 0\) with probability \(1 - \Phi(a_k\theta_i - b_k)\).

For each \(Y_{rep}^{(m)}\), we draw a sample of size \(R\) from the corresponding posterior distribution of ability and item parameters, \(\{\tilde{\theta}^{(r,m)}, \tilde{\xi}^{(r,m)}\}_{r=1}^{R}\). With this we compute a size \(R\) sample
from the posterior distribution of $T_i$ given $Y_{rep}^{(m)}$,

$$T_{i,1} \left( y_{i,rep}^{(m)}, \left( \tilde{\theta}_i^{(1,m)}, \tilde{\xi}_i^{(1,m)} \right) \right), \ldots, T_{i,R} \left( y_{i,rep}^{(m)}, \left( \tilde{\theta}_i^{(R,m)}, \tilde{\xi}_i^{(R,m)} \right) \right),$$  \hspace{1cm} (4.72)

denoted $\{T_{i,l}^{(m)}\}_{l=1}^R$, for $i = 1, \ldots, I$, where $y_{i,rep}^{(m)}$ is the response vector of the $i^{th}$ examinee from $Y_{rep}^{(m)}$.

In using the PPPS method here we want to compare the posterior sample $\{T_i^{(0)}\}$ of the measure $T$ based on $Y_{obs}$ to the posterior samples $\{T_i^{(m)}\}_{r=1}^R$ of $T$ based on the replicated data $Y_{rep}^{(m)}$, for each $i = 1, \ldots, I$. To compare the posterior sample $\{T_i^{(0)}\}$ to the samples $\{T_i^{(m)}\}_{r=1}^R$, we compute the vector of the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles of each sample (denoted $q_i^{(0)}$ for $\{T_i^{(0)}\}$ and $q_i^{(m)}$ for $\{T_i^{(m)}\}_{r=1}^R$) as a summary measure and use these. As a way to compare the quantile vector $q_i^{(0)}$ to $q_i^{(m)}$, $m = 1, \ldots, M$, we compute the Euclidean distances $e_i^{(0)}$ from $q_i^{(0)}$ to $\bar{q}_i$, and $e_i^{(m)}$ from $q_i^{(m)}$ to $\bar{q}_i$, where $\bar{q}_i = (\bar{q}_{i,0.05}, \bar{q}_{i,0.25}, \bar{q}_{i,0.50}, \bar{q}_{i,0.75}, \bar{q}_{i,0.95})^t$ is the mean quantile vector over all $q_i^{(m)}$. The values of $e_i^{(0)}$ and $e_i^{(m)}$, $m = 1, \ldots, M$, are compared by computing the empirical $p$-value,

$$\hat{P}_i(e_i^{(m)} \geq e_i^{(0)}) = \frac{\sum_{m=1}^M I(e_i^{(m)} \geq e_i^{(0)})}{M + 1},$$  \hspace{1cm} (4.73)

where $I(.)$ is an indicator function.

If the proportion of distances $e_i^{(m)}$ greater than $e_i^{(0)}$ is small, this means that $e_i^{(0)}$ is large relative to the other distances. This implies that the quantiles of the posterior sample $\{T_i^{(0)}\}$ are “farther away” from the mean quantiles than those of the posterior samples $\{T_i^{(m)}\}_{r=1}^R$, $m = 1, \ldots, M$. In theory, if the $i^{th}$ examinee’s response vector $y_{i,obs}$ was generated under the model, then we would expect the corresponding posterior sample of $T_i$ to be similar to the others simulated given the replicated data $Y_{rep}^{(m)}$. If the posterior sample of $T_i$ given $Y_{obs}$ is not similar, this casts doubt on $y_{i,obs}$ having been generated under the model. In particular, if $p$-value in (4.73) is smaller than some threshold value of say 0.05, this indicates that the response pattern of examinee $i$ does not fit under the assumed model.
In this process, for examinee $i$, we compare the posterior distribution of the measure $T_i$ based on $Y_{obs}$ to posterior distributions of $T_i$ based on replicated data $Y_{rep}$ under the model. This takes care of the issues mentioned in the last section 4.5.2 about measures having unknown sampling distributions under certain models, and thus leaving the question of what an appropriate reference distribution would be. Here, via simulation, we are referencing the posterior distribution of the measure $T_i$ computed under the observed data to posterior distributions of $T_i$ that would arrive from data produced under the model. So, the person fit measures can be used under this method for different models (1PN, 2PN, 3PN, etc) provided we have a means to simulate replicated data from the prior predictive distribution and then simulate from the posterior.

4.5.3.1 Using the $d_Z$ Measure under the PPPS Method

Here we briefly explain how the $d_Z$ measure can be used under the PPPS method. By simulating posterior draws of the parameters $\tilde{\theta}^{(n)}$ and $\tilde{\xi}^{(n)}$, and latent variables $\tilde{Z}^{(n)}$, $n = 1, \ldots, N$, from their marginal posterior distributions, we would again compute a posterior sample $\{d_{Z_i}^{(n)}\}_{n=1}^{N}$ based on the observed data. Then we would simulate $M$ replicate data sets from under the model, $Y_{rep}^{(m)}$, $m = 1, \ldots, M$. For each $Y_{rep}^{(m)}$, simulate from the resulting posterior a sample of the ability and item parameters, and latent variables. From these posterior samples, we would compute a posterior sample of the measure $\{d_{Z_i}^{(rep,m)}\}_{r=1}^{R}$. Following the necessary steps, we can compute the $p$-value in (4.73) to determine if response pattern of examinee $i$ seems to be aberrant under the assumed model.

4.5.4 Pivotal Discrepancy Measure (PDM) Method

The PDM method is applicable if a measure $T$ is pivotal in that its distribution is invariant when evaluated at the true parameter values, and this distribution is known. These measures can be functions of the data and parameters, $T(Y, (\theta, \xi))$, or parameters alone $T(\theta, \xi)$. If $T$ is pivotal then from Theorem 1 of Johnson (2007) or Lemma 1 of Yuan and Johnson (2011) as
discussed in chapter 2, we can use the distribution of $T$ as a reference distribution to compare posterior values of $T$. In this, we treat a posterior sample of $T$ as a dependent sample from the reference distribution and apply the order statistic bounds to obtain a minimum bound $p$-value, $p_{\text{min}}$ as in (3.23) in section 3.4.4.3.

Due to the restrictions placed on the measure $T$, this method does limit the types of measures that can be used as opposed to the posterior predictive or PPPS methods. However, if we do have measures with known reference distributions then the PDM method is quite easy to implement since all we need is a posterior sample of the parameters to compute a posterior sample of $T$.

Looking at the person fit measures we have, the $W$, $U$, $UB$, $\xi_1$, and $\xi_2$ measures would not be able to be used under the PDM method. Recalling from our discussion of these measures in section 4.3 that the exact distribution they would follow, even when evaluated at the true parameter values, would not be known. However, the $d_Z$ and $L_z$ measures may be used.

### 4.5.4.1 Using the $d_Z$ Measure under the PDM Method

As mentioned in section 4.5.1.4, the measure $d_{Z_i}$ has a known prior distribution which is $\chi^2_K$. This distribution is free of unknown parameters and so $d_Z$ is a pivotal quantity. Based on this, and since $d_Z$ is a function of only unknown parameters, then by Lemma 1 of [Yuan and Johnson (2011)] we can use it under the PDM method. In this, the $\chi^2_K$ prior distribution would serve as the reference distribution. For a posterior sample $\{d_{Z_i}^{(n)}\}_{n=1}^N$ the $p_{\text{min}}$ value would be computed by finding the overall minimum of the order statistic bounds applied to the sample. If the $p_{\text{min}}$ value is less than 0.05, this is evidence that the response pattern of the $i^{th}$ examinee is aberrant under the model.

### 4.5.4.2 Using the $L_z$ Measure under the PDM Method

Here we give some further justification as to why we can use the $L_z$ measure under the PDM method. In section 4.3 it was mentioned that the standardized log-likelihood statistic
$L_z$ has an asymptotic standard normal distribution when it is computed using data generated under the model and the true values of the data generating parameters. This implies that $L_z$ is, at least asymptotically, a pivotal quantity.

The original statement of the $L_z$ measure having an asymptotic standard normal distribution was made while working with the 3PN IRT model (Drasgow et al. 1985, Molenaar and Hoijtink 1990). We are particularly interested in the asymptotic distribution of $L_z$ under the 2PN IRT model since that is the model our research will focus on. We perform a simulation to investigate the asymptotic distribution of $L_z$ under the 2PN IRT model, and also under the 1PN and 3PN IRT models. We only bother with the 1PN and 3PN models because the simulation is easy to perform and if there distributions are close to $N(0,1)$ then this could tell us that $L_z$ could also be used as a person fit measure under the PDM method for these models.

The measure $L_z$, defined in (4.7), has components given in (4.6), (4.8), and (4.9), that all deal with summations over the $K$ items. This means that the sample size that $L_z$ depends on to asymptotically approach a standard normal distribution is the number of items $K$. So, it would be a good idea to see how large the number of items $K$ needs to be for the distribution of $L_z$ to reasonably approximate the $N(0,1)$ distribution.

We will check three test lengths of $K = 20, 40, and 60$ items given to $I = 200$ examinees. We generate an $I \times K$ response data matrix $Y_{obs}$ under the 1PN, 2PN, and 3PN models, and compute the value of $L_{z,i}$ for the $200^{th}$ examinee using the observed responses $y_{obs,i}$ and true values of the ability and item parameters. We repeat this 1,000 times for each of the test lengths to effectively simulate a sample from the distribution of $L_{z,i}$ under each of the three models. The 2PN Model used is the same as given in section 4.1. The 1PN model we use is the same as the 2PN model except we take the item discrimination parameters $a_k \equiv 1$. The 3PN model is also the same except we have an added item guessing parameter $c_k$ which we give a $Beta(2,6)$ prior distribution. The two prior shape parameters are chosen to give $c_k$ a prior mean of 0.25.
We plot the histograms of each computed sample of $L_{z,200}$ in figure 4.6. Each row corresponds to the samples of $L_{z,200}$ computed under a given model, with the top being for the 1PN model, the middle for the 2PN model, and bottom for the 3PN model. Each column of graphs corresponds to a different test length, where the first column has test length of $K = 20$, second column has $K = 40$, and last has $K = 60$. In all graphs, the $N(0,1)$ density curve is plotted for reference.

We see under each model that there is some negative skewness in the samples where the test length was $K = 20$. Looking down the first column of figure 4.6 there appear to be no values in the samples of $L_{z,200}$ above 2 under any of the three models. This skewness tends to be reduced as the test length increases. For a test length of $K = 40$ there is still some negative skewness, especially under the 1PN and 2PN models. In spite of this however, the histograms appear to follow the standard normal curve very closely. Looking at the histograms in the third column, it also appears the distribution of $L_{z,200}$ is very close to standard normal. Since the number of items $K$ used affects the overall time of our MCMC simulation, we will use $K = 40$ in our simulation study at the end of this chapter.

As mentioned in the end of section 3.4.4.3, Johnson (2007) and Yuan and Johnson (2011) allow for the case when a measure may only be a pivotal quantity asymptotically. In this case they recommend that when comparing a posterior sample of the measure with the reference distribution using the computed order statistics, to treat extreme values with caution (Johnson 2007). Yuan and Johnson (2011) recommend discarding a certain percentage of the lowest and highest of the computed order statistics before computing the probability bounds. From this then, $L_z$ can be used under the PDM method.

With the above in mind, under the 2PN IRT model we can simulate a posterior sample of the ability and item parameters via MCMC and use these to compute a posterior sample of the probabilities of success. That is, $\{\hat{P}_{ik}^{(n)}\}_{n=1}^N$, where $\hat{P}_{ik}^{(n)} = \Phi(\tilde{a}_k^{(n)} \tilde{\theta}_i^{(n)} - \tilde{b}_k^{(n)})$, for $i = 1, \ldots, I$ and $k = 1, \ldots, K$. With this and the observed data, we can compute a sample $\{\hat{L}_{z,i}^{(n)}\}_{n=1}^N$ of the measure, for $i = 1, \ldots, I$. We will compute the order statistics of the posterior sample,
Figure 4.6 Samples of 1,000 values of measure $L_z$ computed using simulated data and parameters under the model. Each row gives samples computed under the 1PN, 2PN and 3PN IRT models. Each column corresponds to a test length of $K = 20, 40, \text{ or } 60$. The standard normal density curve is superimposed in all plots for reference.
\( \tilde{L}_{z,i(n)} \) for \( n = 1, \ldots, N \), and discard the top and bottom 1% of the total \( N \) order statistics, as per the discussion above. Then, using the standard normal distribution as the reference distribution we will compute probability bounds on the remaining order statistics and take the overall minimums as the \( p_{\text{min}} \) values (chapter 3). If the value of \( p_{\text{min}} \) for an examinee is less than 0.05, this indicates that their response pattern is aberrant under the model.

### 4.5.5 Discussion Comparing the Bayesian Methods

As mentioned, in Bayesian IRT models the posterior predictive approach is widely used in what research has been done into person fit checking (e.g., Nering and Meijer 1998, Glas and Meijer 2003, and Armstrong et al. 2007). This is due to its ease of implementation, but also that to date there have not been many alternatives. The PPPS method was introduced in Dey et al. (1998) almost fifteen years ago, but has not been looked at much in Bayesian model checking due to its high computational demand. The PDM method is relatively new (Johnson 2007, Yuan and Johnson 2011) and has not been adopted much either.

One of the problems with PP model checking is that it can be conservative (Robins et al. 2000). Yuan and Johnson (2011) performed a simulation study comparing the posterior predictive, PPPS and PDM model checking methods for a Bayesian hierarchical model. The results showed that the PPPS and PDM methods had greater detection rates than the posterior predictive method. The PDM method is definitely cheaper computationally than the PPPS method, but is limited on the number of measures that can be used. However, as we will see, the PPPS method is feasible for smaller amounts of examinees. Also, the main factor against the PPPS method is computation time which hinges on the speed of the computer used. Since the processing speed of computers is constantly being improved the issues of computation time will eventually become moot. Since under the PPPS method we are directly comparing the posterior of a measure based on the observed data to posteriors of the measure based on data generated under the model, this has promise of being a very powerful procedure.
4.6 Example Using Each Bayesian Model Checking Method to Check Person Fit

We conduct an example where we illustrate how each measure can be used to detect examinee lack of fit under each method. We simulate a response data set $Y_{obs}$ with responses of $I = 200$ examinees to $K = 40$ items under the 2PN IRT model from section 4.1. The type of person misfit we will look at detecting is cheating. For our example, cheating is taken to occur when an examinee of low ability answers the more difficult items correctly. The response data $Y$ is simulated such that a total of 5% of the examinees engaged in cheating. Specifically, examinees 10, 30, 50, 70, 90, 110, 130, 150, 170, and 190 were cheaters. The cheating examinees’ responses were simulated by first drawing their abilities from a $U(-2, -0.50)$ distribution (i.e. a low ability level) and assigning a probability of 1 for a correct response to the top 10% most difficult items. For the below examples, we focus on checking the fit of the responses from examinee 50. A threshold value of 0.05 is used for all $p$-values except for the posterior predictive $p$-value corresponding to $L_z$ which is 0.95. The reason for this will be discussed in this next section.

4.6.1 Posterior Predictive Method

We start by using the posterior predictive method. Under this method, we begin by fitting the 2PN model to $Y_{obs}$ using the MCMC Gibbs sampling procedure. We draw a posterior sample of the ability and item parameters $\{\tilde{\theta}^{(l)}, \tilde{a}^{(l)}, \tilde{b}^{(l)}\}_{l=1}^{1000}$ of length 1,000, where $\tilde{\theta}^{(l)} = \{\tilde{\theta}_1^{(l)}, \ldots, \tilde{\theta}_{200}^{(l)}\}^t$, $\tilde{a}^{(l)} = \{\tilde{a}_1^{(l)}, \ldots, \tilde{a}_{40}^{(l)}\}^t$, and $\tilde{b}^{(l)} = \{\tilde{b}_1^{(l)}, \ldots, \tilde{b}_{40}^{(l)}\}^t$, and latent variables $\{Z^{(l)}\}_{l=1}^{1000}$ (for use with the $d_Z$ measure). Using these, we also replicate a sample of $I \times K$ response data matrices, $\tilde{Y}^{(1)}, \ldots, \tilde{Y}^{(1000)}$, from the posterior predictive distribution of $Y$. We simulate $\tilde{Y}^{(l)}$, by simulating each entry $\tilde{y}_{ik}^{(l)}$ as a Bernoulli random variable with probability of success $\tilde{p}_{ik}^{(l)} = \Phi(\tilde{a}_k^{(l)} \tilde{\theta}_i^{(l)} - \tilde{b}_k^{(l)})$ for $i = 1, \ldots, 200$ and $k = 1, \ldots, 40$.

The first measure we will use is $L_z$. Recall from (4.7) that $L_z$ is the standardization of
the log-likelihood measure $L_0$. For examinee 50, we first compute the 1,000 values of $L_{z,50}$ using the vector of observed responses $\mathbf{y}_{50,obs}$ and posterior probabilities of success $\hat{p}_{50k}^{(l)}$. This is given by,

$$L_{z,50}^{(l)} = L_{0,50}^{(l)} - E(L_{0,50}^{(l)}) \sqrt{Var(L_{0,50}^{(l)})}, \quad (4.74)$$

where

$$L_{0,50}^{(l)} = \sum_{k=1}^{K} \left[ y_{50,k} \log(\hat{p}_{50k}^{(l)}) + (1 - y_{50,k}) \log(1 - \hat{p}_{50k}^{(l)}) \right], \quad (4.75)$$

$$E(L_{0,50}^{(l)}) = \sum_{k=1}^{K} \hat{p}_{50k}^{(l)} \log(\hat{p}_{50k}^{(l)}) + (1 - \hat{p}_{50k}^{(l)}) \log(1 - \hat{p}_{50k}^{(l)}) \right], \quad (4.76)$$

and

$$Var(L_{0,50}^{(l)}) = \sum_{k=1}^{K} \hat{p}_{50k}^{(l)} [1 - \hat{p}_{50k}^{(l)}] \left[ \log \frac{\hat{p}_{50k}^{(l)}}{1 - \hat{p}_{50k}^{(l)}} \right]^2, \quad (4.77)$$

for $l = 1, \ldots, 1000$.

We then compute the 1,000 values of $L_z$ using the posterior predictive response vectors $\tilde{\mathbf{y}}_{50}^{(l)} = (\tilde{y}_{1,50}^{(l)}, \ldots, \tilde{y}_{K,50}^{(l)})^t$ and posterior probabilities $\hat{p}_{50k}^{(l)}$. In this we compute the posterior predictive values of the log-likelihood measure $L_{0,50}$ as,

$$\tilde{L}_{0,50}^{(l)} = \sum_{k=1}^{K} \left[ \tilde{y}_{50,k} \log(\hat{p}_{50,k}^{(l)}) + (1 - \tilde{y}_{50,k}) \log(1 - \hat{p}_{50,k}^{(l)}) \right], \quad (4.78)$$

for $l = 1, \ldots, 1000$ and then use the expected values and variances from (4.76) and (4.77), respectively, to compute a posterior predictive sample of $L_{z,50}$. This is given by,

$$\tilde{L}_{z,50}^{(l)} = \frac{\tilde{L}_{0,50}^{(l)} - E(L_{0,50}^{(l)})}{\sqrt{Var(L_{0,50}^{(l)})}}, \quad (4.79)$$
for \( l = 1, \ldots, 1000 \). The posterior predictive \( p \)-value for examinee 50 is then computed as,

\[
P_{PP(L_z)}^{(50)} = \frac{\sum_{l=1}^{1000} I[\tilde{L}_{z,50}^{(l)} \geq L_{z,50}^{(l)}]}{1000} = 0.991 .
\] (4.80)

Recall from chapter 3 that posterior predictive \( p \)-values close to 0 or 1 are taken as evidence of misfit. When the \( L_z \) measure is computed under the PP method using response data that is aberrant under the model, the \( p \)-values tend to be larger (i.e. closer to 1) rather than smaller (i.e. closer to 0). For this reason, we choose a threshold value of 0.95 for posterior predictive \( p \)-values of \( L_z \). Based on a threshold value of 0.95, this \( p \)-value indicates that the item responses of examinee 50 are not fitted well by the model.

To compute the posterior predictive \( p \)-value for \( d_{Z,50} \) we first compute the 1,000 posterior values of \( d_{Z,50}^{(l)} \), given the observed data. These are computed by taking the posterior sample of Bayesian latent variables for examinee 50, \( Z_{50}^{(l)} = (Z_{50,1}^{(l)}, \ldots, Z_{50,40}^{(l)}) \) for \( l = 1, \ldots, 1000 \), along with the simulated posterior values of the ability and item parameters \( \tilde{\theta}_{50}^{(l)}, \tilde{a}^{(l)}, \) and \( \tilde{b}^{(l)} \). Using these, we compute a posterior sample of Bayesian latent residuals \( e_{50,k}^{(l)} = Z_{50,k}^{(l)} - \tilde{a}^{(l)} \tilde{\theta}^{(l)}_{50}_k + \tilde{b}^{(l)}_k \) for \( k = 1, \ldots, 40 \) and \( l = 1, \ldots, 1000 \). These values are then used to compute the posterior sample of \( d_{Z,50} \) given by,

\[
d_{Z,50}^{(l)} = \sum_{k=1}^{40} (e_{50,k}^{(l)})^2 , \quad \text{for} \ l = 1, \ldots, 1000 .
\] (4.81)

Then the posterior predictive sample of \( d_{Z,50} \) is computed using the sample of posterior predictive data \( \tilde{Y}^{(l)} \) and posterior parameters \( \{\tilde{\theta}^{(l)}, \tilde{a}^{(l)}, \tilde{b}^{(l)}\} \), \( l = 1, \ldots, 1000 \). With the posterior predictive response data for examinee 50, \( \tilde{y}_{50}^{(l)} = (\tilde{y}_{50,1}^{(l)}, \ldots, \tilde{y}_{50,40}^{(l)}) \), and posterior parameter values \( \tilde{\theta}_{50}^{(l)}, \tilde{a}_k^{(l)}, \) and \( \tilde{b}_k^{(l)} \), \( k = 1, \ldots, 40 \), we simulate a posterior predictive Bayesian latent variable \( \tilde{Z}_{50,k}^{(l)} \sim TN(0,\infty)(\tilde{a}_k^{(l)} \tilde{\theta}_{50}^{(l)} - \tilde{b}_k^{(l)}, 1) \), if \( \tilde{y}_{50,k}^{(l)} = 1 \) and \( \tilde{Z}_{50,k}^{(l)} \sim TN(-\infty,0)(\tilde{a}_k^{(l)} \tilde{\theta}_{50}^{(l)} - \tilde{b}_k^{(l)}, 1) \), if \( \tilde{y}_{50,k}^{(l)} = 0 \). From this posterior predictive sample of Bayesian latent variables we then
compute a posterior predictive sample of Bayesian latent residuals \( \tilde{\epsilon}_{50,k} = \tilde{Z}_{50,k}^{(l)} - \tilde{a}_k^{(l)} \tilde{\theta}_{50}^{(l)} + \tilde{b}_k^{(l)} \)

and finally a posterior predictive sample for the measure \( d_{Z,50} \) given by,

\[
\tilde{d}_{Z,50}^{(l)} = \sum_{k=1}^{40} (\tilde{\epsilon}_{50,k}^{(l)})^2,
\]

for \( l = 1, \ldots, 1000 \).

The posterior predictive \( p \)-value is then computed as,

\[
\hat{p}_{PP(d_Z)}^{(50)} = \frac{\sum_{l=1}^{1000} I[\tilde{d}_{Z,50}^{(l)} \geq d_{Z,50}^{(l)}]}{1000} = 0.174.
\]

The \( d_Z \) measure does not detect the lack of fit in the response pattern of examinee 50. As we will also see in our simulation study, this measure does not show to be very powerful under the PP method.

Next we will compute the posterior predictive \( p \)-values for the residual based measures, \( U, W, \) and \( UB \) for examinee 50. We first compute the posterior samples of the measures using the observed responses \( y_{50,obs} \) and posterior probabilities of success \( \tilde{p}_{50k}^{(l)} \). These are given by,

\[
U_{50}^{(l)} = \frac{1}{K} \sum_{k=1}^{K} \frac{(y_{50,k} - \tilde{p}_{50,k})^2}{\tilde{v}_{50,k}},
\]

and

\[
W_{50}^{(l)} = \frac{\sum_{k=1}^{K} (y_{50,k} - \tilde{p}_{50,k})^2}{\sum_{k=1}^{K} \tilde{v}_{50,k}},
\]

for \( l = 1, \ldots, 1000 \) where \( \tilde{v}_{50,k} = \tilde{p}_{50,k}(1 - \tilde{p}_{50,k}) \).

For computing \( UB_{50}^{(l)} \), we first rank the items from easiest to most difficult just like in the classical example in section 4.4. In that example we ranked the items according to the mle’s of their difficulty parameters. Here we use the sample means of the simulated posterior difficulty parameters. We then split the items into \( S = 2 \) subsets of size 20 each, denoted \( A_s, s = 1, 2 \). Subset \( A_1 \) contains the bottom 20 easiest items and \( A_2 \) contains the top 20
hardest items. We then compute the 1,000 values of $UB_{50}^{(l)}$,

$$UB_{50}^{(l)} = \frac{1}{2 - 1} \left[ \sum_{j \in A_s} (y_{50,j} - \tilde{p}_{50,j}) \right]^2 \sum_{j \in A_s} \tilde{v}_{50,j},$$

(4.86)

for $l = 1, \ldots, 1000$.

Now, using the replicated posterior predictive responses and posterior probabilities we compute the posterior predictive sample of $U_{50}$, $W_{50}$, and $UB_{50}$,

$$\tilde{U}_{50}^{(l)} = \frac{1}{K} \sum_{k=1}^{K} \left( \tilde{y}_{50,k} - \tilde{p}_{50,k} \right)^2 \tilde{v}_{50,k},$$

(4.87)

$$\tilde{W}_{50}^{(l)} = \frac{\sum_{k=1}^{K} (\tilde{y}_{50,k} - \tilde{p}_{50,k})^2}{\sum_{k=1}^{K} \tilde{v}_{50,k}},$$

(4.88)

and

$$\tilde{UB}_{50}^{(l)} = \frac{1}{2 - 1} \left[ \sum_{j \in A_s} (\tilde{y}_{50,j} - \tilde{p}_{50,j}) \right]^2 \sum_{j \in A_s} \tilde{v}_{50,j},$$

(4.89)

for $l = 1, \ldots, 1000$. The posterior predictive $p$-values are then computed as,

$$\hat{p}_{PP(U)}^{(50)} = \frac{\sum_{l=1}^{1000} I[\tilde{U}_{50}^{(l)} \geq U_{50}^{(l)}]}{1000} = 0.015,$$

(4.90)

$$\hat{p}_{PP(W)}^{(50)} = \frac{\sum_{l=1}^{1000} I[\tilde{W}_{50}^{(l)} \geq W_{50}^{(l)}]}{1000} = 0.029,$$

(4.91)
\[ \hat{p}_{PP(UB)}^{(50)} = \frac{\sum_{l=1}^{1000} I(\tilde{U}B_{50}^{(l)} \geq UB_{50})}{1000} = 0.0, \] (4.92)

which are all less than 0.05.

For finding the posterior predictive p-values of the two remaining measures \( \xi_1 \) and \( \xi_2 \) we also start by computing the sample of values \( \tilde{\xi}_{1,50}^{(l)} \) and \( \tilde{\xi}_{2,50}^{(l)} \) for \( l = 1, \ldots, 1000 \), using the observed responses of examinee 50 and the sampled posterior probabilities. Referring to (4.35) and (4.37) we compute,

\[ \tilde{\xi}_{1,50}^{(l)} = \frac{\sum_{k=1}^{K}(\tilde{G}_k^{(l)} - \tilde{G}_k^{(t)})\tilde{\xi}_{50,k}^{(l)} - y_{50,k}^{(t)}}{\left(\sum_{k=1}^{K} \tilde{v}_{50,k}^{(l)}(\tilde{G}_k^{(l)} - \tilde{G}_k^{(t)})^2\right)^{1/2}}. \] (4.93)

and

\[ \tilde{\xi}_{2,50}^{(l)} = \frac{\sum_{k=1}^{K}(\tilde{p}_{50,k}^{(l)} - \tilde{T}_{50})\tilde{\xi}_{50,k}^{(l)} - y_{50,k}^{(t)}}{\left(\sum_{k=1}^{K} \tilde{v}_{50,k}^{(l)}(\tilde{p}_{50,k}^{(l)} - \tilde{T}_{50})^2\right)^{1/2}}, \] (4.94)

where for each posterior draw \( l \), \( \tilde{G}_k^{(l)} = \frac{1}{T} \sum_{i=1}^{T} \tilde{p}_{ik}^{(l)}, \tilde{G}_k^{(t)} = \frac{1}{K} \sum_{k=1}^{K} \tilde{G}_k^{(t)}, \) and \( \tilde{T}_{50} = \frac{1}{K} \sum_{k=1}^{K} \tilde{p}_{50,k}^{(t)} \).

Then we compute the posterior predictive values of \( \tilde{\xi}_{1,50}^{(l)} \) and \( \tilde{\xi}_{2,50}^{(l)} \) for \( l = 1, \ldots, 1000 \). These are given as,

\[ \tilde{\xi}_{1,50}^{(l)} = \frac{\sum_{k=1}^{K}(\tilde{G}_k^{(l)} - \tilde{G}_k^{(t)})\tilde{\xi}_{50,k}^{(l)} - y_{50,k}^{(t)}}{\left(\sum_{k=1}^{K} \tilde{v}_{50,k}^{(l)}(\tilde{G}_k^{(l)} - \tilde{G}_k^{(t)})^2\right)^{1/2}}, \] (4.95)

and

\[ \tilde{\xi}_{2,50}^{(l)} = \frac{\sum_{k=1}^{K}(\tilde{p}_{50,k}^{(l)} - \tilde{T}_{50})\tilde{\xi}_{50,k}^{(l)} - y_{50,k}^{(t)}}{\left(\sum_{k=1}^{K} \tilde{v}_{50,k}^{(l)}(\tilde{p}_{50,k}^{(l)} - \tilde{T}_{50})^2\right)^{1/2}}. \] (4.96)
The posterior predictive p-values are then found to be,

\[ \hat{p}_{PP}(\xi_1) = \frac{\sum_{l=1}^{1000} I[\tilde{\xi}_{1,50}^{(l)} \geq \xi_{1,50}]}{1000} = 0.0, \]

and

\[ \hat{p}_{PP}(\xi_2) = \frac{\sum_{l=1}^{1000} I[\tilde{\xi}_{2,50}^{(l)} \geq \xi_{2,50}]}{1000} = 0.327. \]

The p-value for \( \xi_{1,50} \) is 0 and that for \( \xi_{2,50} \) is 0.327. So, the measure \( \xi_1 \) identifies the aberrancy in examinee 50’s responses and \( \xi_2 \) does not. As we will see later in our simulation study, the measure \( \xi_2 \) does not show to be a powerful measure in detecting cheating under the PP method.

### 4.6.2 Prior Predictive Posterior Simulation Method

In using the PPPS method, our first step is the same as that of the posterior predictive method. That is, we fit the 2PN model to the simulated data \( Y_{obs} \) using the MCMC procedure, and draw a posterior sample of the parameters \( \{\tilde{\theta}^{(l)}, \tilde{a}^{(l)}, \tilde{b}^{(l)}\}^{1000}_{l=1} \), and latent variables \( \{Z^{(l)}\}^{1000}_{l=1} \). We then compute the 1,000 posterior values of the measures \( L_{z,50}, d_{z,50}, U_{50}, W_{50}, UB_{50}, \xi_{1,50}, \) and \( \xi_{2,50} \) given the simulated observed responses \( y_{50,obs} \) as in (4.74), (4.81), (4.84), (4.85), (4.86), (4.93), and (4.94), respectively, of the previous section 4.6.1. We then compute the 1 × 5 vector of the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles of the 1,000 posterior values for each measure, denoted \( q_{\theta_{z,50}}^{(0)}, q_{d_{z,50}}^{(0)}, q_{y_{50}}^{(0)}, q_{w_{50}}^{(0)}, q_{ub_{50}}^{(0)}, q_{\xi_{1,50}}^{(0)}, \) and \( q_{\xi_{2,50}}^{(0)} \).

Next, we replicate 100 data sets \( Y_{rep}^{(m)}, m = 1, \ldots, 100, \) from the prior predictive distribution using the steps indicated in section 4.5.3. That is, for each \( m \) we simulate the parameters from their prior distributions: \( \theta_i^{(m)} \sim N(0,1), i = 1, \ldots, 200, a_k^{(m)} \sim TN(0,\infty)(1,50) \)
and \(b_k^{(m)} \sim N(0,1)\), for \(k = 1, \ldots, 40\). Then, we simulate dichotomous responses \(y_{ik,rep}^{(m)}\), where \(y_{ik,rep}^{(m)} = 1\) with probability \(\Phi(a_k^{(m)}\theta_i^{(m)} - b_k^{(m)})\) and \(y_{ik,rep}^{(m)} = 0\) with probability \(1 - \Phi(a_k^{(m)}\theta_i^{(m)} - b_k^{(m)})\).

For each \(Y_{rep}^{(m)}\), we then fit the 2PN model using the MCMC procedure and draw a sample of size \(R = 1,000\) from the corresponding posterior distribution of ability and item parameters \(\{\tilde{\theta}^{(r,m)}, \tilde{a}^{(r,m)}, \tilde{b}^{(r,m)}\}_{r=1}^{1000}\), where \(\tilde{\theta}^{(r,m)} = (\tilde{\theta}_1^{(r,m)}, \ldots, \tilde{\theta}_{200}^{(r,m)})\), \(\tilde{a}^{(r,m)} = (\tilde{a}_1^{(r,m)}, \ldots, \tilde{a}_{40}^{(r,m)})\), \(\tilde{b}^{(r,m)} = (\tilde{b}_1^{(r,m)}, \ldots, \tilde{b}_{40}^{(r,m)})\), and and latent variables \(\{Z^{(r,m)}\}_{r=1}^{1000}\). Given these, the following steps are done to compute a p-value for each measure, except \(d_Z\), to determine the fit the observed response vector \(y_{50,obs}\). For the measure \(d_Z\) the there is only a slight difference in step (1) below, and will be discussed shortly. For the sake of explanation we will describe the following steps using just the \(L_z\) measure.

We start with step 1: for each \(m\) we compute the 1,000 values of \(L_z,50\) using the vector of replicated responses \(y_{50,rep}^{(m)}\) and posterior parameters \(\tilde{\theta}_{50}^{(r,m)}, \tilde{a}^{(r,m)}, \text{ and } \tilde{b}^{(r,m)}, r = 1, \ldots, 1000.\) This is given by,

\[
\begin{align*}
L_{z,50}^{(r,m)} &= \frac{L_{0,50}^{(r,m)} - E(L_{0,50}^{(r,m)})}{\sqrt{Var(L_{0,50}^{(r,m)})}},
\end{align*}
\]

where

\[
L_{0,50}^{(r,m)} = \sum_{k=1}^{K} \left[ y_{50,rep}^{(m)} \log(\tilde{p}_5^{(r,m)}) + (1 - y_{50,rep}^{(m)}) \log(1 - \tilde{p}_5^{(r,m)}) \right],
\]

\[
E(L_{0,50}^{(r,m)}) = \sum_{k=1}^{K} \left[ \tilde{p}_5^{(r,m)} \log(\tilde{p}_5^{(r,m)}) + (1 - \tilde{p}_5^{(r,m)}) \log(1 - \tilde{p}_5^{(r,m)}) \right],
\]

and

\[
Var(L_{0,50}^{(r,m)}) = \sum_{k=1}^{K} \tilde{p}_5^{(r,m)} [1 - \tilde{p}_5^{(r,m)}] \left[ \log \frac{\tilde{p}_5^{(r,m)}}{1 - \tilde{p}_5^{(r,m)}} \right]^2,
\]

for \(r = 1, \ldots, 1000\), where \(\tilde{p}_5^{(r,m)} = \Phi(\tilde{\theta}_5^{(r,m)} - \tilde{b}_5^{(r,m)})\). The same is done for the other measures where \(U_{50}^{(m)}, W_{50}^{(m)}, UB_{50}^{(m)}, \xi_{1,50}, \text{ and } \xi_{2,50}\) are computed as in \((4.84), (4.85), (4.86), (4.93), \text{ and } (4.94)\), respectively, using \(y_{50,rep}^{(m)}\) and posterior probabilities \(\tilde{p}_5^{(r,m)} = \Phi(\tilde{\theta}_5^{(r,m)} - \tilde{b}_5^{(r,m)})\).
The 1,000 posterior values of $d_{Z,50}$, given $y_{50,rep}^{(m)}$, are computed using the posterior parameters $\tilde{\theta}_{50}^{(r,m)}$, $\tilde{a}^{(r,m)}$, $\tilde{b}^{(r,m)}$, and latent variables $Z^{(r,m)}_{50} = (Z^{(r,m)}_{50,1}, \ldots, Z^{(r,m)}_{50,100})$, $r = 1, \ldots, 1000$. We compute the latent residuals $\epsilon_{50,k}^{(r,m)} = Z^{(r,m)}_{50,k} - \tilde{a}_k^{(r,m)} \tilde{\theta}_{50}^{(r,m)} + \tilde{b}_k^{(r,m)}$ and then the sample of the measure $d_{Z,50}^{(r,m)} = \sum_{k=1}^{K} (\epsilon_{50,k}^{(r,m)})^2$, $r = 1, \ldots, 1000$.

Step 2: we compute the $1 \times 5$ vector of the 0.05, 0.25, 0.50, 0.75, and 0.95 quantiles over the 1,000 posterior values $L_{z,50}^{(r,m)}$, denoted $q_{L_{z,50}}^{(m)}$. Step 3: we compute the mean vector of the 100 quantile vectors $q_{L_{z,50}}^{(m)}$, $m = 1, \ldots, 100$, denoted $\bar{q}_{L_{z,50}}$. Step 4: compute the Euclidean distance, $e_{L_{z,50}}^{(0,m)}$ between each of the 100 quantile vectors $q_{L_{z,50}}^{(m)}$, and the mean quantile vector $\bar{q}_{L_{z,50}}$. Step 5: we compute the Euclidean distance $e_{L_{z,50}}^{(0)}$ between the quantile vector $q_{L_{z,50}}^{(0)}$ and $\bar{q}_{L_{z,50}}$. Finally, an empirical p-value is computed as,

$$p_{PPPS(L_z)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{L_{z,50}}^{(m)} \geq e_{L_{z,50}}^{(0)})}{101} = 0.010,$$

where $I(.)$ is an indicator function.

Similarly in (2) through (5) above, each of $q_{d_{z,50}}^{(m)}$, $q_{U_{50}}^{(m)}$, $q_{W_{50}}^{(m)}$, $q_{UB_{50}}^{(m)}$, $q_{\xi_{1,50}}^{(m)}$, and $q_{\xi_{2,50}}^{(m)}$ are computed along with the mean quantile vectors $\bar{q}_{d_{z,50}}$, $\bar{q}_{U_{50}}$, $\bar{q}_{W_{50}}$, $\bar{q}_{UB_{50}}$, $\bar{q}_{\xi_{1,50}}$, and $\bar{q}_{\xi_{2,50}}$, and corresponding Euclidean distances $e_{d_{z,50}}^{(m)}$, $e_{U_{50}}^{(m)}$, $e_{W_{50}}^{(m)}$, $e_{UB_{50}}^{(m)}$, $e_{\xi_{1,50}}^{(m)}$, and $e_{\xi_{2,50}}^{(m)}$. We have the computed empirical p-values for the rest of the measures, starting with $d_{Z,50}$ as,

$$p_{PPPS(d_z)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{d_{z,50}}^{(m)} \geq e_{d_{z,50}}^{(0)})}{101} = 0.010,$$
\[ \hat{p}_{\text{PPPS}(U)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{U50}^{(m)} \geq e_{U50}^{(0)})}{101} = 0.040, \]  

\[ \hat{p}_{\text{PPPS}(W)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{W50}^{(m)} \geq e_{W50}^{(0)})}{101} = 0.010, \]  

\[ \hat{p}_{\text{PPPS}(\xi)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{\xi2.50}^{(m)} \geq e_{\xi2.50}^{(0)})}{101} = 0.010, \]  

\[ \hat{p}_{\text{PPPS}(\xi2)}^{(50)} = \frac{\sum_{m=1}^{100} I(e_{\xi2.50}^{(m)} \geq e_{\xi2.50}^{(0)})}{101} = 0.010. \]  

The \( p \)-values corresponding to each of the measures is less than 0.05, so each measure is able to detect the aberrant response pattern of examinee 50 under the PPPS method.
4.6.3 Example with PDM Method

For the PDM method, as with the previous two methods, we fit the 2PN model to $Y_{obs}$ using the MCMC procedure and draw off a posterior sample of 1,000 values of the ability and item parameters $\{\theta^{(l)}, \tilde{a}^{(l)}, \tilde{b}^{(l)}\}_{l=1}^{1000}$, and latent variables $\{Z^{(l)}\}_{l=1}^{1000}$. We then compute 1,000 posterior values of the measures $L_{z,50}^{(l)}$ and $d_{Z,50}^{(l)}$, $l = 1, \ldots, 1000$, given the observed response vector $y_{50,obs}$ as in (4.74) and (4.81).

As discussed previously in sections 4.5.4.1 and 4.5.4.2 the measures $L_{z,i}$ and $d_{Z,i}$ for $i = 1, \ldots, 200$, are pivotal quantities. When evaluated using the data generating parameters, the measure $L_{z,i}$ has an asymptotic standard normal distribution and $d_{Z,i}$ has a $\chi^2_{40}$ distribution. So, we will use the standard normal distribution as a reference distribution to compare the posterior sample of $L_{z,50}$. Likewise, the $\chi^2_{40}$ distribution will be the reference distribution for the posterior sample of $d_{Z,50}$.

Recall, in the PDM method we compare the posterior sample of the measure to the reference distribution by way of the order statistics. In this, we order the 1,000 posterior values of $L_{z,50}^{(l)}$ and $d_{Z,50}^{(l)}$ to get the sample order statistics, $L_{z,50(l)}$ and $d_{Z,50(l)}$ for $l = 1, \ldots, 1000$, respectively. Then we compute the minimum of the order statistic bounds discussed from section 3.4.4.3 to get a minimum $p$-value for each measure, $p_{min(L_z)}^{(50)}$ and $p_{min(d_Z)}^{(50)}$. These are computed as,

$$p_{min(L_z)}^{(50)} = \min \left\{ \min \left\{ 1, \frac{1000[1 - \Phi(L_{z,50(l)})]}{1000 - l - 1} \right\}_{l=1}^{1000}, \min \left\{ \frac{1000\Phi(L_{z,50(l)})}{l}, 1 \right\}_{l=1}^{1000} \right\}$$

and

$$p_{min(d_Z)}^{(50)} = \min \left\{ \min \left\{ 1, \frac{1000[1 - F(d_{Z,50(l)})]}{1000 - l - 1} \right\}_{l=1}^{1000}, \min \left\{ \frac{1000F(d_{Z,50(l)})}{l}, 1 \right\}_{l=1}^{1000} \right\}$$

$$= 0.003$$

and

$$= 0.011.$$
where $\Phi(.)$ is the standard normal cdf and $F(.)$ represents the cdf of the $\chi^2_{40}$ distribution. Using the rule of thumb of Yuan and Johnson (2011) discussed in chapter 3, $p_{\text{min}}$ values less than 0.05 indicate a strong lack of fit of the response pattern with the model. We see both $d_Z$ and $L_z$ correctly identify the aberrant response pattern of examinee 50.

4.7 Person Fit Simulation Study with the 2PN IRT Model

In this section we will discuss a simulation study that was done to investigate the performance of the person fit measures in detecting response violations under a 2PN IRT model. In the study we will look at the type I error rate produced by each measure, under each of the three model checking methods: posterior predictive, PDM, and PPPS. Then we compare the detection rates for the different measures within each model checking method to get an idea of which measures are most powerful. For each violation type, we then compare the average detection rates between each method. This will allow us to examine under which method a measure is more powerful.

For this study we use data sets of $I = 200$ examinees and $K = 40$ items. The number of items is selected to be 40 because that is considered a sufficient test length for the $L_z$ measure to have a $N(0, 1)$ distribution when evaluated with the data and data generating parameters (see section 4.5.4.2). Two-hundred examinees is low in comparison to other IRT person fit research (e.g. Smith 1985, Glas and Meijer 2003), but we are using this amount due to the computational demand of the PPPS method. Glas and Meijer (2003), for example, showed that the detection rates of the measures $L_z$, $\xi_1$, $\xi_2$, $UB$, and $W$ used under the posterior predictive method tended to increase as the number of examinees increased. However, as we shall see, the measures used under the PPPS method have respectable detection rates for each type of response violation we will look at.
4.7.1 Description of the Study

4.7.1.1 Types of Response Pattern Violations Used

As mentioned, in person fit we are determining how well an examinee’s response pattern fits the model. There are various ways examinees can develop aberrant response patterns. We focus on the following as part of our study.

Types of person misfit:

1. **Guessing**: The definition of guessing is not unanimously agreed upon in the field of test theory. Here we take the version from Wright and Stone (1979), a person of low ability tends to get difficult items correct. This makes the assumption that the examinee would try to answer items of low to moderate difficulty honestly and only “guess” on the items that were too difficult for them.

2. **Sleeping**: This is described in Wright and Stone (1979) also. Sleeping or fumbling, is where an examinee of high ability tends to get easy items incorrect. Since an examinee of high ability should be expected to get easier items correct, the reason for the incorrect responses is assumed to be due to some clumsiness or not paying attention.

3. **Random Responding**: In this, examinees blindly respond to test items in such a way that they have a constant probability of success. For example, if a student were taking a multiple-choice test with, say four possible choices A, B, C, or D, picking an answer at random would yield a probability of a correct response of 25% (Karabatsos 2003). Intuitively, it sounds like this particular type of responding should be classified as guessing, but this is how it is referred to in the literature, so we make the separate distinction here.

4. **Cheating**: Like guessing, this can be looked at in many different ways, but here we assume a “cheater” is someone of low ability who answers the more difficult items on the test correctly (Karabatsos 2003). Since an examinee of low ability should be
expected to get harder items incorrect, the correct responses are assumed to be due to
the examinee obtaining the right answers dishonestly.

4.7.1.2 2PN Model Specification and Simulation Specifics

We use the same Bayesian 2PN IRT model given in the beginning of the chapter in section
4.1 which includes a data augmentation step. We restate the model here for the reader’s
convenience,

\[
y_{ik} = \begin{cases} 
1, & \text{if } Z_{ik} > 0 \\
0, & \text{if } Z_{ik} \leq 0 
\end{cases} \tag{4.110}
\]

\[
Z_{ik} \sim N(a_k \theta_i - b_k, 1) \tag{4.111}
\]

\[
\theta_i \sim N(0, 1), \tag{4.112}
\]

\[
a_k \sim TN(0, \infty)(1, .50), \tag{4.113}
\]

\[
b_k \sim N(0, 1), \tag{4.114}
\]

for \(i = 1, \ldots, I\) and \(k = 1, \ldots, K\). All the measures: \(d_z, L_z, U, W, UB, \xi_1,\) and \(\xi_2\) were used
to check the fit of the response patterns in the simulated data sets under the PP and PPPS
methods. Since only the sum of squared Bayesian latent residuals \(d_z\) and the standardized
log-likelihood \(L_z\) qualified as pivotal quantities, these were the only ones used in the PDM
method.

For each of the four types of response violations, and for each of the three model checking
procedures, 100 data sets were simulated and the fit of each examinee’s response pattern
was checked. All posterior samples were drawn using the MCMC Gibbs sampling routine of
Albert (1992) for the 2PN model outlined in section 2.5.2. The chains were run to a length
of 7,000 with a burn-in period of 4,000. The remaining chains of 3,000 were thinned by one-
third to give posterior samples of length 1,000. For each iteration under the PPPS method,
we simulated 100 separate data sets \( Y_{rep} \) from the prior predictive distribution, computed the posterior distribution for each, and evaluated the person fit measures on posterior samples of size 1,000.

Significant \( p \)-values were determined as follows. Under the PP method, a \( p \)-value less than 0.05 was taken to imply significant deviation of an examinee’s response pattern from that expected under the model for all measures except \( L_z \), in which a \( p \)-value of 0.95 was used. Under the PPPS method, a \( p \)-value less than 0.05 was deemed as significant for all measures. For the PDM method, the rule-of-thumb was applied and the measure was considered to have detected an aberrancy in an examinee’s response pattern if \( p_{min} \) was less than 0.05.

### 4.7.2 Simulation Study Results: Type I Error Rates

We investigated the type I error rate for each of the measures evaluated under each model checking method. For each of the three model checking methods, we simulate 100 data sets under the 2PN model. For each data set we checked the fit of each examinee using the above person fit measures and record the number of significant \( p \)-values. This means that for each measure under say, the PDM method, we get a vector of length 200 where the \( i^{th} \) entry gives the number of significant \( p \)-values out of 100 for the \( i^{th} \) examinee.

In table 4.6 we list the average number of significant \( p \)-values over all 200 examinees for each measure, and under each method. Since there were a total of 100 simulations under each method, we can look at this number as the average number of times out of 100 an examinee’s response pattern was indicated as misfitting, i.e. the type I error rate.

We see that the average number of times an examinee was indicated as having an aberrant response pattern, when in fact it was derived under the model, was quite low for the measures evaluated under the PP method. Under the PPPS method we see the average type I error rate among the examinees are larger than under the PP method, being closer to around 5%. Some measures have a mean type I error rate slightly greater than 5% under this method, such as \( \xi_1 \) or \( \xi_2 \) with 5.2% and 5.3%, respectively. Under the PDM method, the average type
I error rate of $d_Z$ and $L_z$ was about 0.9% and 3.9%, respectively.

Overall, the measures seem to show a good type I error rate under all methods in that they do not incorrectly identify an examinee as misfitting much more than the threshold level of 5%. The $d_Z$ measure tended to have the lowest rates under the PP and PDM methods, but as we will see in the next section this measure also tended to not be very powerful under these methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>$d_Z$</th>
<th>$L_z$</th>
<th>$U$</th>
<th>$W$</th>
<th>$UB$</th>
<th>$\xi_1$</th>
<th>$\xi_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>0.0</td>
<td>2.3</td>
<td>1.6</td>
<td>2.3</td>
<td>1.7</td>
<td>2.8</td>
<td>3.3</td>
</tr>
<tr>
<td>PPPS</td>
<td>5.1</td>
<td>5.1</td>
<td>4.9</td>
<td>4.9</td>
<td>4.6</td>
<td>5.2</td>
<td>5.3</td>
</tr>
<tr>
<td>PDM</td>
<td>0.9</td>
<td>3.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6 Average type I error rates over the 200 examinees for each person fit measure evaluated under each method: PP, PPPS, and PDM. Under the PDM method, the type I error rates for $d_Z$ and $L_z$ are given applying the rule-of-thumb where rejection is taken if $p_{min} < 0.05$.

4.7.3 Simulation Study Results: Detection Rates

In this section we present the results of the simulation study to investigate the detection rates of the person fit measures under the 2PN IRT model. First we will explain how we simulated each of the four types of response violations: guessing, sleeping, random response, and cheating. We simulate data sets under each of the four types of response violations where 10, or 5%, of the examinees produced aberrant response patterns. In the simulation studies done by Karabatsos (2003) it was found that detection rates were highest when the aberrant responders were at 5%, and the rates decreased for higher percentages of examinees with aberrant responses. So, we chose 5% to try to produce the largest detection rates possible to compare the measures. For simplicity, the ten examinees producing response violations were examinees: 10, 30, 50, 70, 90, 110, 130, 150, 170, and 190.
The four response violations were simulated in the following ways. To simulate guessing, the 10 violating examinees’ abilities were drawn from a $U(-2, -0.5)$ distribution and their responses to items with difficulty parameter $b_k > 1$, i.e. difficult items, were simulated with a probability of correct response being set at 0.70. For all items with difficulty parameter $b_k \leq 1$ the responses were generated with a probability of success $p_{ik} = \Phi(a_k\theta_i - b_k)$. To simulate sleeping, the examinees’ high abilities were simulated from a $U(0.5, 2)$ distribution. Then, for items with difficulty $b_k < -1$, i.e. easy items, their responses were generated with probability of a correct response being 0.30. Random responding was generated by giving each item a 0.25 probability of being correctly answered. Cheating examinees’ responses were simulated by first drawing their abilities from a $U(-2, -0.5)$ distribution, and then generating responses to the top 10% most difficult items with a success probability of 1.

The study was conducted in the following way. Under each of the three model checking methods, four observed data sets corresponding to each response violation type were simulated. For each data set, we draw a posterior sample of the item and ability parameters, and latent variables using the MCMC Gibbs sampling procedure. We then evaluate the person fit measures for each of the 200 examinees and compute the 200 corresponding $p$-values. For each examinee, we record if the $p$-value was significant.

This is repeated a total of 100 times. For each method, we then compile the total number of significant $p$-values for each examinee according to violation type and measure used. For example, each measure used under the PP method will have four vectors of length 200 corresponding to it, one for each violation type. The $i^{th}$ entry of each vector will give the total number of significant $p$-values out of 100 for the $i^{th}$ examinee.

In the next two sections we discuss the results of the person fit measures in detecting the four response violations under the three model checking methods. From a practical standpoint, when we have different methods and measures at our disposal a logical question we might ask is “what is the best method to use?”, or “what is the best measure to use?” to detect a given response violation. However, it is not that simple. There are different factors
to consider: the method being employed, the measure being used, and the violation type being detected. As we will see, certain measures will perform better than others depending on the model checking method it is used under and the response violation being imposed.

We will analyze the results from two perspectives. First, we will look at the performance of the measures under each model checking method separately. This is to help us understand, that if one given method was to be employed, which measures perform better than others in detecting the violations. Secondly, for each response violation type we will focus on comparing how each measure performs between the three methods.

4.7.3.1 Performance of Person Fit Measures Within Each Method

We would first like to compare the performance of the person fit measures within each of the three model checking methods. Under each of the four response violation types, we take the mean of the detection rates for each measure over the 10 violating examinees. Tables 4.7, 4.8, and 4.9 give the results obtained under the PP, PDM, and PPPS model checking methods, respectively. In each table, the measures’ average detection rates are listed for each violation type, along with the respective estimated standard errors. For each model checking method, we also plot the mean detection rates in trellis dot plots which combine the results for each type of violation. These are given in figures 4.7, 4.8, and 4.9 We now discuss the results under each method, starting with the PP method followed by the PDM and PPPS methods.

4.7.3.1.1 PP Method

From table 4.7 and figure 4.7 we see a pretty consistent pattern for the measures used under the PP method. The two best measures tended to be \( \xi_1 \) and \( L_z \). Following these, the residual based measures \( U, W, \) and \( UB \), in varying orders, had the next best detection rates. The \( \xi_2 \) and \( d_z \) measures did not do very well in detecting the response violations, the exception being for Random Response, in which \( \xi_2 \) had the highest average detection rate.
Average Detection Rates Under Posterior Predictive Method

<table>
<thead>
<tr>
<th>Measures</th>
<th>Guessing</th>
<th>Sleeping</th>
<th>Random Response</th>
<th>Cheating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.E.</td>
<td>Mean</td>
<td>S.E.</td>
</tr>
<tr>
<td>$d_Z$</td>
<td>5.2</td>
<td>2.0</td>
<td>5.3</td>
<td>1.6</td>
</tr>
<tr>
<td>$L_z$</td>
<td>60.7</td>
<td>3.1</td>
<td>70.1</td>
<td>2.8</td>
</tr>
<tr>
<td>$U$</td>
<td>55.7</td>
<td>3.4</td>
<td>66.8</td>
<td>3.5</td>
</tr>
<tr>
<td>$W$</td>
<td>53.2</td>
<td>3.2</td>
<td>61.4</td>
<td>3.7</td>
</tr>
<tr>
<td>$UB$</td>
<td>55.6</td>
<td>4.5</td>
<td>60.5</td>
<td>3.5</td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>72.8</td>
<td>2.6</td>
<td>77.4</td>
<td>4.2</td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>39.0</td>
<td>3.9</td>
<td>44.4</td>
<td>5.6</td>
</tr>
</tbody>
</table>

Table 4.7 Mean detection rates of the person fit measures evaluated using the posterior predictive method. For each person fit measure, the mean is taken over the detection rates corresponding to the 10 violating examinees. This is given for each violation type. Also, for each measure the estimated standard error of the mean detection rate is given for each violation type.

We will discuss the results here for each measure under each of the four response violations.

1. **Guessing**

   For detecting Guessing under the PP method the best measure was $\xi_1$ with a mean detection rate around 73%. The next best was $L_z$ with a mean rate of about 60% followed by $U$, $UB$, and $W$ with rates between 53% and 56%. The measures $\xi_2$ and $d_Z$ had the lowest mean rates at 39% and 5.2%, respectively.

2. **Sleeping**

   For Sleeping, the $\xi_1$ measure has the highest average detection rates at around 77% followed by $L_z$ at around 70%. The measures $U$, $W$, and $UB$ had around 67%, 61%, and 60%, respectively. The $\xi_2$ and $d_Z$ measures had the lowest mean rates at about 44% and 5%, respectively.

3. **Random Response**

   Here, we see that the $\xi_2$ measure has the highest rates at around 74%, followed by $\xi_1$ at about 71% and thirdly $L_z$ at about 66%. The average detection rates for the $W$, $U$,
Figure 4.7 A dot plot for each response violation type shows the average detection rates of each person fit measure calculated while performing the posterior predictive model checking method.
and $UB$ measures were lower at around 65%, 61%, and 53%, respectively. The lowest average rates being were from $d_Z$ at around 15%.

4. Cheating

For this violation, $\xi_1$, $L_z$, had the highest average rates at approximately 97% and 92%, respectively. Next, $W$ had a mean rate of about 87% followed by $UB$ at around 80% and $U$ being much lower at around 64%. The lowest average detection rates corresponded to $\xi_2$ and $d_Z$ at about 39% and 0.3% respectively.

Suppose we wanted to identify the best measures to use under the PP method for detecting the four response violations. We have already seen that the measures with the largest mean rates for detecting Guessing, Sleeping, and Cheating were $\xi_1$, followed by $L_z$. So, when using the PP method we may be apt to recommend using $\xi_1$ for detecting each of the three violation types. However, if we look at the estimated standard errors (s.e.) of the mean rates for these measures, one may wish to use either $\xi_1$ or $L_z$.

For example, in detecting Sleeping we see that the $\xi_1$ measure has a mean rate of 77.4% and an estimated s.e. of 4.2 while the $L_z$ measure has a mean detection rate of 70.1% and an estimated s.e. of 2.8. The measure $\xi_1$ has a higher detection rate, but also a higher estimated s.e. Even though $L_z$ has a lower detection rate, a researcher may view the smaller estimated s.e. as an indication that $L_z$ is more reliable.

In detecting Guessing and Cheating, we can say that $\xi_1$ appears to be the best measure since it has the highest mean detection rate and smallest estimated s.e. The exception of course being for the $d_Z$ measure, but its detection rates are low enough to exclude it from consideration. For detecting Random Responding, $\xi_2$ has the highest mean detection rate and lowest estimated s.e. of 2.9 (again, excluding $d_Z$). However, to its credit $\xi_1$ would be the second best since it has the second highest detection rate and second lowest estimated s.e.

4.7.3.1.2 PDM Method
Table 4.8 and figure 4.8 show the average detection rates for the $d_Z$ and $L_z$ measures under the PDM method. Under the PDM method, rejection (and hence, detection of person misfit) was taken if $p_{min} < .05$. In table 4.8, each row gives the mean number of times in 100 simulations that a significant $p$-value (i.e. $p_{min} < .05$) was obtained for the violating examinees. Under each violation type in figure 4.8, we also plot these two values for each measure.

<table>
<thead>
<tr>
<th>Violation Type</th>
<th>d_Z</th>
<th>L_z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guessing</td>
<td>Mean</td>
<td>S.E.</td>
</tr>
<tr>
<td>39.8</td>
<td>4.4</td>
<td>40.9</td>
</tr>
<tr>
<td>80.6</td>
<td>2.7</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Table 4.8 Mean detection rates of the person fit measures $d_Z$ and $L_z$ evaluated using the pivotal discrepancy measure (PDM) method. For the two pivotal person fit measures, the mean number of times in 100 simulations that a significant $p$-value (i.e. $p_{min} < .05$) was obtained for the 10 violating examinees is given along with each mean’s standard error. This is given for each violation type.

A dot plot for each response violation type shows the mean detection rates, i.e. the mean number of times in 100 simulations that a significant $p$-value (i.e. $p_{min} < .05$) was obtained for the 10 violating examinees of each person fit measure used under the PDM model checking method. The two measures used were the $d_Z$ and $L_z$ measures since these were the only measures that were pivotal quantities.

1. **Guessing**

For detecting Guessing, the $L_z$ measure has the larger mean detection rate at 80.6%. The $d_Z$ measure had a mean rate of only about 40%.

2. **Sleeping**

For detecting Sleeping, the $L_z$ measure also has a respectable mean detection rate of 79.3%. The $d_Z$ measure had a mean rate of only about 41%.
Figure 4.8 A dot plot for each response violation type shows the mean detection rates, i.e. the mean number of times in 100 simulations that a significant $p$-value (i.e. $p_{min} < .05$) was obtained for the 10 violating examinees of each person fit measure used under the PDM model checking method. The two measures used were the $d_Z$ and $L_z$ measures since these were the only measures that were pivotal quantities.

3. **Random Response**

The $L_z$ measure did not do quite as well in detecting Random Responding, having a mean rate of about 65%. The $d_Z$ measure had a little better mean rate of about 48%, but still not very good.

4. **Cheating**

The $L_z$ measure performed the best in detecting Cheating in the violating examinees with a mean detection rate of 93.8%. The $d_Z$ measure performed worst in detecting Cheating with a mean rate of about 32%.
If we want to decide which is the better measure to use under the PDM method the decision is quite clear. The $L_z$ measure has a larger mean detection rate over the $d_Z$ measure by a considerable margin for all four violation types. Except for the Sleeping violation, the mean rates of the $L_z$ measure also have the lower estimated standard errors.

### 4.7.3.1.3 PPPS Method

The average detection rates of the 10 violating examinees for all person fit measures evaluated under the PPPS method are plotted in figure 4.9 and listed in table 4.9. The performances under each violation type are discussed below.

<table>
<thead>
<tr>
<th>Measures</th>
<th>Guessing Mean</th>
<th>Guessing S.E.</th>
<th>Sleeping Mean</th>
<th>Sleeping S.E.</th>
<th>Random Response Mean</th>
<th>Random Response S.E.</th>
<th>Cheating Mean</th>
<th>Cheating S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_Z$</td>
<td>80.4</td>
<td>2.6</td>
<td>82.8</td>
<td>3.6</td>
<td>65.1</td>
<td>5.1</td>
<td>94.0</td>
<td>2.6</td>
</tr>
<tr>
<td>$L_z$</td>
<td>78.0</td>
<td>2.7</td>
<td>80.4</td>
<td>3.8</td>
<td>65.2</td>
<td>4.8</td>
<td>93.7</td>
<td>2.5</td>
</tr>
<tr>
<td>$U$</td>
<td>61.1</td>
<td>3.5</td>
<td>63.3</td>
<td>4.7</td>
<td>53.3</td>
<td>5.8</td>
<td>42.7</td>
<td>5.6</td>
</tr>
<tr>
<td>$W$</td>
<td>70.5</td>
<td>4.2</td>
<td>72.6</td>
<td>3.1</td>
<td>83.4</td>
<td>2.1</td>
<td>86.5</td>
<td>3.9</td>
</tr>
<tr>
<td>$UB$</td>
<td>71.1</td>
<td>3.2</td>
<td>72.6</td>
<td>3.5</td>
<td>64.9</td>
<td>3.9</td>
<td>92.1</td>
<td>2.5</td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>80.2</td>
<td>2.4</td>
<td>81.6</td>
<td>3.8</td>
<td>76.0</td>
<td>4.2</td>
<td>93.6</td>
<td>2.6</td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>50.4</td>
<td>5.3</td>
<td>51.2</td>
<td>4.4</td>
<td>84.5</td>
<td>2.6</td>
<td>48.0</td>
<td>2.8</td>
</tr>
</tbody>
</table>

Table 4.9 Mean detection rates of the person fit measures evaluated using the prior posterior predictive simulation (PPPS) method. For each person fit measure, the mean is taken over the detection rates corresponding to the 10 violating examinees. This is given for each violation type. Also, for each measure the estimated standard error of the mean detection rate is given for each violation type.

1. **Guessing**

For Guessing, the $d_Z$, $\xi_1$, and $L_z$ measures had the highest rates with 80.4%, 80.2%, and 78%, respectively. The next best group of measures would be $UB$ and $W$ with mean detection rates of about 71%. The measure $U$ had a lower mean rate with about 61% and $\xi_2$ had the lowest rate with close to 50%.
2. Sleeping

In detecting Sleeping, the measures with the highest mean detection rates were again $d_Z$, $\xi_1$, and $L_z$ with 82.8%, 81.6%, and 80.4%, respectively. The measures $W$ and $UB$ both had the same mean rate of 72.6%. The next highest mean rate was for $U$ with about 63% and $\xi_2$ had the lowest with close to 51%.

3. Random Response

The best two measures for detecting Random Responding were $\xi_2$ and $W$. These measures had average rates of about 85% and 83%, respectively. The next best measure was $\xi_1$ with a mean rate of 76%. The rates for measures $d_Z$, $L_z$, and $UB$ were very close, all being about 65%. The measure with the lowest mean detection rate for Random
Responding was $U$ with about 53%.

4. **Cheating**

In detecting Cheating, almost all measures used under the PPPS method had very high average rates. The highest average rate was obtained by the $d_Z$ measure at 94% with $L_z$ and $\xi_1$ very close with 93.7% and 93.6%, respectively, and $UB$ measures were greater than 97%. The measures $W$ and $UB$ had mean rates of about 92% and 87%, respectively. The lowest average rates were obtained from $U$ and $\xi_2$, both being below 50%.

For detecting Guessing, Sleeping, and Cheating under the PPPS method there is a pattern in the mean detection rates of the measures. Looking at table 4.9 and figure 4.9 we can see the $d_Z$ measure has the highest mean rate for these three violation types followed by $\xi_1$ and $L_z$ having rates only slightly less. The group of measures with the next lowest average rates is $W$ and $UB$. The measures $U$ and $\xi_2$ have the lowest mean detection rates. For detecting Random Responding using the PPPS method however, we see that $\xi_2$ performs best followed very closely by $W$.

Suppose we now wanted to identify the best measures to use under the PPPS method for detecting the four response violations. We have already seen that the measures with the largest mean rates for detecting Guessing, Sleeping, and Cheating were $d_Z$, $\xi_1$, and $L_z$, with $d_Z$ being the highest by a narrow margin. If we look at the estimated standard errors (s.e. ) of the mean rates for each of these measures, one may wish to use either $\xi_1$ or $L_z$.

For instance, in detecting Guessing we see that the $d_Z$ measure has a mean rate of 80.4% and an estimated s.e. of 2.6 while the $\xi_1$ measure has a mean detection rate of 80.2% and an estimated s.e. of 2.4. Due to the detection rates being so close one may wish to use $\xi_1$ based on the slightly lower estimated s.e. Similarly in detecting Cheating, $L_z$ has a slightly lower mean rate than $d_Z$, but also a slightly lower s.e.
In detecting Sleeping, we can say that $d_Z$ appears to be the best measure since it has the highest mean detection rate and smallest estimated s.e. For detecting Random Responding, $\xi_2$ has the highest mean detection rate and a low estimated s.e. of 2.6. However, although we see that $W$ has a slightly lower mean rate of 83.4%, it has a lower s.e. of 2.1. So, this may seem a more desirable measure.

### 4.7.3.2 Performance of Person Fit Measures Between Each Method

In the previous section we focused on how the measures performed in detecting the different response violations within each model checking method: PP, PPPS, and PDM. Now, for each of the four response violations, we will focus on comparing the individual measures’ performance between each method. The purpose of this is to see under which method a given measure will perform best for a particular response violation.

We take the information from tables 4.7, 4.8, and 4.9 and reorganize it. A table is created for each response violation type, in which we list the mean detection rates of the person fit measures obtained under the three model checking methods. We put this information in tables 4.10, 4.11, 4.12, and 4.13 corresponding to Guessing, Sleeping, Random Responding, and Cheating, respectively. In each table, we also list the estimated standard errors of the mean detection rates under each method. Again, $d_Z$ and $L_z$ are the only measures to have rates listed under the PDM method since these were the only pivotal measures to be used.

For a better visual comparison we also plot the mean detection rates in parallel bar charts in figures 4.10 and 4.11. The mean rates of each measure in detecting Guessing and Sleeping are plotted in figure 4.10, and the rates in detecting Random Responding and Cheating are plotted in figure 4.11. Next, we will discuss the results under each response violation type.

In table 4.10 we can see under which method each measure performed better for detecting Guessing. The $d_Z$ measure does best when used under the PPPS method with a mean detection rate of about 80% versus about 40% under the PDM method and 5% under the PP method. The highest mean rate for $L_z$ is under the PDM method at 80.6% followed by
78% under the PPPS method. Both of these are improvements over the performance under the PP method which yielded a mean rate of about 61%. Looking at the estimated s.e.’s under the PPPS and PDM methods we see they are the same at 2.7. This would indicate that $L_z$ performs best under the PPPS method.

<table>
<thead>
<tr>
<th>Measures</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_Z$</td>
<td>5.2</td>
<td>80.4</td>
<td>39.8</td>
<td>2.0</td>
<td>2.8</td>
<td>4.4</td>
</tr>
<tr>
<td>$L_z$</td>
<td>60.7</td>
<td>78.0</td>
<td>80.6</td>
<td>3.1</td>
<td>2.7</td>
<td>2.7</td>
</tr>
<tr>
<td>$U$</td>
<td>55.7</td>
<td>61.1</td>
<td></td>
<td>3.4</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>$W$</td>
<td>53.2</td>
<td>70.5</td>
<td></td>
<td>3.2</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>$UB$</td>
<td>55.6</td>
<td>71.1</td>
<td></td>
<td>4.5</td>
<td>3.2</td>
<td></td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>72.8</td>
<td>80.2</td>
<td></td>
<td>2.6</td>
<td>2.4</td>
<td></td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>39.0</td>
<td>50.4</td>
<td></td>
<td>3.9</td>
<td>5.3</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.10 Average rates of the person fit measures detecting a Guessing response violation over the 10 violating examinees. The average rates are given under each of the three model checking methods: posterior predictive (PP), prior posterior predictive simulation (PPPS), and pivotal discrepancy measure (PDM). The estimated standard errors are also given.

For measures $U$, $W$, $UB$, $\xi_1$, and $\xi_2$ the only comparison is between the PP and PPPS methods. We see that in detecting Guessing, all these measures had higher average detection rates under the PPPS method. However, some of the estimated s.e.’s were also increased.

The measure $U$ has the smallest increase in mean detection rate of 5.4% from 55.7% under the PP method to 61.1% under the PPPS method. The estimated s.e. only increased slightly from 3.4 to 3.5 though. So, it appears that $U$ performs better in detecting Guessing under the PPPS method although it is still not a particularly powerful measure for this response violation type.

There is a good increase of about 17% from the PP to the PPPS method for $W$ to detect an average of about 70% of the Guessing response violations. The estimated s.e. did increase from 3.2 to 4.2 however. The $UB$ measure’s mean rate increased from 55.6% under the PP method to 71.1% under the PPPS method and also had a decrease of estimated s.e. from 4.5
to 3.2. This is a good indication that $UB$ detects Guessing better under the PPPS method than under the PP method.

As mentioned, both $\xi_1$ and $\xi_2$ also have higher mean detection rates under the PPPS method. The $\xi_2$ measure has a low mean rate under the PP method of about 39% with an increase to about 50% under the PPPS method. The $\xi_1$ measure has a respectable mean detection rate of 72.8% under the PP method and about 8% more than that under the PPPS method. We also see that the estimated s.e. under the PPPS method is slightly less. This indicates that the PPPS method is the better method to use $\xi_1$ to detect Guessing.

From Table 4.11 we can see that the highest mean rate for detecting Sleeping for the $d_z$ measure was under the PPPS method. This was about 83% compared to about 40% and 5% under the PDM and PP methods, respectively. The $L_z$ measure has mean detection rates of 80.4% under the PPPS method and 79.3% under the PDM method. The mean rate was about 10% less under the PP method, so the performance under both the PPPS and PDM methods seem to be an improvement. The estimated s.e. under the PPPS method is 3.8 versus 4.5 under the PDM method. This would indicate that $L_z$ performs best under the PPPS method in detecting Sleeping.

The measure $U$ did better under the PP than the PPPS method in terms of mean detection rate and estimated s.e. The measures $W$ and $UB$ both had mean detection rates of 72.6% under PPPS method, which was about 11% higher than that under the PP method. The estimated s.e. for $W$ was lower under the PPPS method, indicating that this is the better method to use. For $UB$, the estimated s.e. of the mean detection rate was the same under both the PP and PPPS methods, at 3.5. Due to the higher mean detection rate though, it is clear that $UB$ performs better under the PPPS method in detecting Sleeping.

The $\xi_1$ measure had a slightly higher mean rate of about 82% under the PPPS method and a smaller estimated s.e. This suggests that $\xi_1$ is best used under the PPPS method to detect Sleeping. There is an approximate 7% increase in the mean rate for $\xi_2$ under the
Average Detection Rates for Sleeping Examinees

<table>
<thead>
<tr>
<th>Measures</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_Z$</td>
<td>5.3</td>
<td>82.8</td>
<td>40.9</td>
<td>1.6</td>
<td>3.6</td>
<td>3.0</td>
</tr>
<tr>
<td>$L_z$</td>
<td>70.1</td>
<td>80.4</td>
<td>79.3</td>
<td>2.8</td>
<td>3.8</td>
<td>4.5</td>
</tr>
<tr>
<td>$U$</td>
<td>66.8</td>
<td>63.3</td>
<td></td>
<td>3.5</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>$W$</td>
<td>61.4</td>
<td>72.6</td>
<td></td>
<td>3.7</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td>$UB$</td>
<td>60.5</td>
<td>72.6</td>
<td></td>
<td>3.5</td>
<td>3.5</td>
<td></td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>77.4</td>
<td>81.6</td>
<td></td>
<td>4.2</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>44.4</td>
<td>51.2</td>
<td></td>
<td>5.6</td>
<td>4.4</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.11 Average rates of the person fit measures detecting Sleeping over the 10 violating examinees. The average rates are given under each of the three model checking methods: posterior predictive (PP), prior predictive posterior simulation (PPPS), and pivotal discrepancy measure (PDM). The estimated standard errors are also given.

PPPS versus the PP method, and a smaller estimated s.e. From either method though, $\xi_2$ does not seem to be a very effective measure in detecting Sleeping.

The mean rates of the person fit measures in detecting Random Responding for the 10 violating examinees is compiled into table 4.12. The $d_Z$ measure performed the best under the PPPS method with a mean rate of about 65% and about 48% under the PDM method. Under the PP method, $d_Z$ was only able to detect an average of about 15%. The $L_z$ measure performed just about equally well under all three model checking methods. The highest mean rate being 65.7% under the PP method the lowest being 65.1% under the PDM method. The estimated s.e. under the PDM method was lower than those under the PP or PPPS methods by more than one though. This might imply that the PDM method is the best method of the three for using $L_z$ to detect Random Responding.

The measure $U$ had a higher average detection rate under the PP method with about 61% versus about 53% under the PPPS method. The estimated s.e. was also lower under the PP method by 1.7. Although $U$ does not perform exceedingly well, the PP method seems to be the better method for using it to detect Random Responding.

For the next four measures, $W$, $UB$, $\xi_1$, and $\xi_2$ we see that they all performed better under the PPPS rather than the PP method. The mean rate for $W$ under the PPPS method was a respectable 83.7% versus only about 65% under the PP method. The estimated s.e. of
Figure 4.10 For the *Guessing* and *Sleeping* response violations we plot parallel bar charts showing the average detection rates for each measure evaluated under each appropriate model checking method: PP, PPPS, and PDM. Since only the $L_z$ and $d_Z$ measures were pivotal quantities, only they have bars for the PDM method.

<table>
<thead>
<tr>
<th>Measures</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
<th>PP</th>
<th>PPPS</th>
<th>PDM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_Z$</td>
<td>15.2</td>
<td>65.1</td>
<td>48.4</td>
<td>1.4</td>
<td>5.1</td>
<td>4.4</td>
</tr>
<tr>
<td>$L_z$</td>
<td>65.7</td>
<td>65.2</td>
<td>65.1</td>
<td>4.9</td>
<td>4.8</td>
<td>3.4</td>
</tr>
<tr>
<td>$U$</td>
<td>61.4</td>
<td>53.3</td>
<td></td>
<td>4.1</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td>$W$</td>
<td>64.7</td>
<td>83.4</td>
<td></td>
<td>4.2</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td>$UB$</td>
<td>53.1</td>
<td>64.9</td>
<td></td>
<td>5.5</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>70.5</td>
<td>76.0</td>
<td></td>
<td>4.0</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>73.9</td>
<td>84.5</td>
<td></td>
<td>2.9</td>
<td>2.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.12 Average rates of the person fit measures detecting Random Responding over the 10 violating examinees. The average rates are given under each of the three model checking methods: posterior predictive (PP), prior predictive posterior simulation (PPPS), and pivotal discrepancy measure (PDM). The estimated standard errors are also given.
2.1 under the PPPS method was also lower by almost 2. The performance of $UB$ in detecting Random Responding was not very good under either method, but the larger mean rate of about 65% was under the PPPS method. Its estimated s.e. of 3.9 was also lower by 1.6 than that under the PP method.

The measure $\xi_1$ had a good mean detection rate of 76% under the PPPS method, but still performed decently under the PP method with an average detection rate of about 71%. The estimated s.e. was about the same under both methods at about 4, being slightly lower for the PP method. The higher detection rate then would indicate the PPPS method is better to use $\xi_1$ under.

For $\xi_2$, we see that it had the overall highest mean rates in detecting Random Responding under both the PP and PPPS methods. It had a mean rate of about 74% under the PP method, but had a mean rate of about 11% higher under the PPPS method. The estimated s.e. was slightly lower under the PPPS method at 2.6. This shows that $\xi_2$ is best used under the PPPS method.

The measures’ mean rates for detecting the Cheating response violation are given in table 4.13. We see that most all the measures do very well in detecting this type of response violation under any method used. An exception being the $d_Z$ measure which did very well under the PPPS method with a mean rate of 94%, but had much lower rates of 31.5% and 0.3% under the PDM and PP methods, respectively. The $L_z$ measure had close mean detection rates under the PP, PPPS, and PDM methods at 92.2%, 93.7%, and 93.8%, respectively. The lowest estimated s.e. for $L_z$ was under the PP method at 2.1 followed by 2.5 under the PPPS method and 2.8 under the PDM method. This makes it a little harder to pick the best method to use $L_z$ under. Since the mean rates are so close it might be the researcher’s choice to use the method that $L_z$ has the least variation in detecting Cheating, i.e. the PP method.

The $U$, $W$, and $\xi_1$ measures all had higher mean detection rates under the PP method. The measure $U$ did not perform that well in detecting Cheating under the PP method with
Average Detection Rates for Cheating Examinees

<table>
<thead>
<tr>
<th>Measures</th>
<th>Mean Detection Rates</th>
<th>Estimated Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PP</td>
<td>PPPS</td>
</tr>
<tr>
<td>$d'_Z$</td>
<td>0.3</td>
<td>94.0</td>
</tr>
<tr>
<td>$L_z$</td>
<td>92.2</td>
<td>93.7</td>
</tr>
<tr>
<td>$U$</td>
<td>64.4</td>
<td>42.7</td>
</tr>
<tr>
<td>$W$</td>
<td>87.2</td>
<td>86.5</td>
</tr>
<tr>
<td>$UB$</td>
<td>80.6</td>
<td>92.1</td>
</tr>
<tr>
<td>$\xi_1$</td>
<td>96.7</td>
<td>93.6</td>
</tr>
<tr>
<td>$\xi_2$</td>
<td>39.3</td>
<td>48.0</td>
</tr>
</tbody>
</table>

Table 4.13 Average rates of the person fit measures detecting Cheating over the 10 violating examinees. The average rates are given under each of the three model checking methods: posterior predictive (PP), prior predictive posterior simulation (PPPS), and pivotal discrepancy measure (PDM). The estimated standard errors are also given.

A mean rate of 64.4%, but this was over 20% higher than that under the PPPS method. The mean rate for $W$ under the PP method was 87.2% which was less than 1% higher than that under the PPPS method. So, $W$ seems to detect Cheating well under either method, but the estimated s.e. under the PP method is 2.7 versus 3.9 under the PPPS method. Likewise, $\xi_1$ has a mean rate of about 97% under the PP method and about 94% under the PPPS method. Both are very good, but we see that the estimated s.e. is lower under the PP method at 1.6.

The mean rate of measure $UB$ was respectable under the PP method at about 81%, but was over 11% higher under the PPPS method. The estimated s.e. was lower under the PP method at 2.3 versus 2.5 under the PPPS method. These are close enough that the larger mean detection rate would indicate that $UB$ performs better under the PPPS method.

The measure $\xi_2$ did not perform well in detecting Cheating under either method with mean rates of 39.3% and 48% under the PP and PPPS methods, respectively. The estimated s.e. under the PPPS method is 2.8 which is 1.5 less than that under the PP method. So, $\xi_2$ does perform better under the PPPS method, but does not appear to be an overall effective measure for detecting Cheating.
4.7.3.3 Overall Best Measure/Method Pairing for Each Response Violation

So far we have discussed how each measure performs within each model checking method in section 4.7.3.1 and also how they perform between each method in section 4.7.3.2. As a conclusion, it may be helpful to identify which measure and method pair worked the best in detection for each response violation. That is which measure, under which method, had the overall highest mean detection rate? In doing this we refer back to tables 4.10, 4.11, 4.12 and 4.13 for each response violation and look to see what the highest overall mean detection rate was.
For detecting Guessing, we see from table 4.10 that the overall largest mean detection rate was 80.4% for the $d_Z$ measure under the PPPS method. We note a close second was the $\xi_1$ measure under the PPPS method with a mean rate of 80.2%. The estimated s.e.’s were very close at 2.8 and 2.4, respectively. For detecting Sleeping, we see in table 4.11 that again the $d_Z$ measure under the PPPS method has the highest average rate at 82.8%. In table 4.12 we see that highest mean rate for detecting Random Responding is 84.5%. This is obtained by the $\xi_2$ measure under the PPPS method. We also see that this has one of the smallest estimated s.e.’s as well. Finally, in table 4.13 we see the largest mean rate for detecting Cheating is 96.7%. This is obtained by the $\xi_1$ measure under the PP method.

4.7.4 Application To A Real Data Set

We will now apply these procedures to a real data set. We will use response data from the 2008 Bowling Green State University (BGSU) Mathematics placement test. Before incoming students can register for any Mathematics courses at BGSU they must take a placement test, the outcome of which will determine which level of courses the students will be allowed to register for. Three versions of the test are given, Form A, Form B, and Form C. Our data set consists of the responses of 1022 examinees to the 35 items on the Form A placement test.

We started by fitting the data under the Bayesian 2PN IRT model described in section 4.7.1.2 using the MCMC Gibbs sampling algorithm. We drew samples \( \{\tilde{\theta}^{(l)}, \tilde{a}^{(l)}, \tilde{b}^{(l)}\}_{l=1}^{1000} \) of the ability and item parameters and latent variables \( \{Z^{(l)}\}_{l=1}^{1000} \), from their respective marginal posterior distributions. As in previous discussion, we are trying to identify examinees whose response patterns deviate from what should occur under the assumed model. In doing this, the person fit measure we choose to use is the $L_z$ measure. This is because it can be used under all three model checking methods, and aside from the Random Response violation, it showed some very respectable detection rates in the simulation study under each method.

The $L_z$ measure was evaluated for each examinee under the PP, PPPS and PDM methods...
via the same procedures outlined in the respective examples of section 4.6. The $L_z$ measure indicated 3, 41, and 22 examinees under the PP, PPPS and PDM methods respectively, whose responses were aberrant under the model. Among those examinees indicated to have aberrant responses, 21 were indicated by both the PPPS and PDM methods, and only 1 by all 3. Now, in the simulation study we simulated responses with particular violations and saw how well the measures detected these under the different methods. The question here now is, “For a given set of data in which an examinee(s) response pattern is indicated as misfitting, how do we know what type of violation has occurred?”

A useful tool in Bayesian IRT when analyzing examinee response patterns is the Bayesian latent residual plot. We discussed these in section 4.5.1.2 and used them to identify the outlying responses from a guessing examinee. Toribio (2006) also applied these in an IRT setting to analyze aberrant examinee response patterns to see if guessing could be determined. From our MCMC simulation, we have the posterior samples of the ability and item parameters, and latent variables. From these, we can easily compute posterior samples of the latent residuals, $\tilde{\epsilon}_{ik}^{(l)} = Z_{ik}^{(l)} - \tilde{a}_k^{(l)} \tilde{\theta}_i^{(l)} + \tilde{b}_k^{(l)}$. Then, for a given examinee $i$, we can plot the posterior sample of latent residuals for each item $k = 1, \ldots, 35$, similar to what was done in figure 4.2 in section 4.5.1.3. Specifically, we can construct a 90% probability interval plot using a box plot extending from the 5th to the 95th percentiles.

Figure 4.12 shows the 90% latent residual probability interval box plots for 6 examinees from the 21 identified under the PPPS and PDM methods. In each plot, the box plots are plotted from left to right in order of least difficult item to most difficult. This was based on the difficulty parameter estimates taken as the means of the posterior samples of each $b_k$, $k = 1, \ldots, K$. As discussed in section 4.5.1.2, when the posterior distribution of a latent residual is far from zero, this indicates an outlying response under the model. We use this in the plots below to indicate which responses to items are predicted under the model and which are aberrant by looking to see if the 90% probability interval contains zero. Since an interest of ours is to see which intervals do not include zero, a horizontal line at zero is also
plotted for reference.

In these plots we can see which items were responded to correctly/incorrectly by seeing if the median, marked by a dash in the box of each box plot, is above/below the zero line. So, we can see each examinee’s response pattern, and also which responses were not predicted under the model. This may allow us to at least make an educated guess as to what might have happened with each examinee, i.e. guessing, random responding, etc.

For example, examinees 150 and 546 show almost a constant flip flop between getting answers correct/incorrect throughout the entire range of item difficulty level. This clearly violates what we said a Guttman pattern should look like at the beginning of the chapter and so it is no surprise why these examinees’ response patterns were indicated as aberrant under the model. This constant change may be due to some type of random responding.

If we look at the plot for examinee 14, we see that only six items were answered correctly. Among the items that were answered correctly, three were very difficult. The posterior ability estimate for examinee 14 was $-2.25$, which raises questions as to how the examinee was able to get these difficult questions correct. We can not say for sure, but this may be due to guessing or cheating.

Examinee 101 exhibits a response pattern that might be explained by guessing. The examinee’s ability estimate was $-0.94$. For having a lower ability this examinee’s response pattern seems to follow an appropriate pattern. That is, as the items become more difficult the amount of correct responses reduces. We do see that the first three items were answered incorrectly, which is unexpected as indicated by the box plots far away from zero in the upper right of figure 4.12. However, we also see that three of the most difficult items were unexpectedly answered correctly. This is because these items may have been guessed at.
Figure 4.12 90% probability interval plots of Bayesian latent residuals for 6 examinees identified with misfitting response patterns by the $L_2$ person fit measure evaluated using the PP, PPPS, and PDM methods.
CHAPTER 5

BAYESIAN MODEL CHECKING FOR 2PN TESTLET MODELS

5.1 Introduction

In this chapter we will discuss a new type of model known as a *testlet* model \( \text{[Bradlow et al., 1999; Wainer et al., 2007]} \) and its differences from the IRT models we have looked at so far. Briefly, a *testlet* \( \text{[Wainer and Kiely, 1987]} \) is a group of items on an exam that are based on a common stimulus. For example, a word problem on a mathematics exam may have more than one question associated with it. All items relating to this one passage would make up a testlet. A consequence of having these groups of items related to common stimuli is that there may be some dependence among their responses within each testlet. If this happens, then the assumption of *local independence* (see section 2.2.4 and 4.2) made in the usual IRT models is no longer valid.

\( \text{[Bradlow et al., 1999]} \) introduced a Bayesian testlet model based on Bayesian IRT models to handle this testlet effect. As we will see, if response data \( Y \) is generated under a testlet model there are issues that arise in parameter estimation if we choose to fit an IRT model to \( Y \). So, there is a need in developing model checking procedures that would indicate a lack of fit when an IRT model is fit to data that is generated under a testlet model.

There has been some research done on this so far by \( \text{[Sinharay and Johnson, 2003]} \) which looked at using discrepancy measures based on the sample biserial correlation coefficients under a posterior predictive (PP) approach. The purpose of this chapter will be to further investigate using measures based on the biserial correlations under the PP method and will also investigate three new measures we propose that can be used under the PPPS method. These will be examined and compared in a simulation study at the end of the chapter.
We will begin the chapter by introducing the testlet model and discussing in detail the issues that arise when an IRT model is fitted to testlet data. Then we will introduce the biserial and point biserial correlations and discuss how these have been used as model checking measures for IRT models under the PP method. We will then discuss how they can be used as measures for checking the fit of IRT models to testlet data. Lastly, we will introduce the three measures we propose to use under the PPPS method and our motivation for choosing these before looking at the simulation study.

5.2 Testlet Model

Before introducing the model, we shall specify the data structure. Like before we have an \( I \times K \) response matrix \( Y \), but the items are grouped into \( L \) (\( 1 \leq L \leq K \)) mutually exclusive and exhaustive testlets. A testlet (Wainer and Kiely 1987) is a group of items that are based on a common stimulus. Examples of testlets can be found in tests of reading comprehension where a set of questions might all be related to a particular reading passage, or a mathematics test where a set of questions might be asked about a certain data set. The testlet of item \( k \) is denoted by \( d(k) \), \( d(k) \in \{1, \ldots, L\} \), and the size of each testlet by \( n_l \). A testlet with \( n_l = 1 \) is considered an independent item.

If testlets are present in a test, then an individual’s responses to those items in a particular testlet may be viewed as being dependent (even if conditioning on the ability and item parameters). That is, the assumption of local independence that was made for IRT models (see chapter 1) will not hold. We may view item responses between different testlets as being independent, but responses within testlets will have a dependence structure (Bradlow et al. 1999).

5.2.1 Two-Parameter Testlet Model

Bradlow et al. (1999) introduced a version of the Bayesian two-parameter probit model
that incorporated a random effects parameter to deal with the dependence in individuals’ item responses within testlets. This model is also given a brief discussion in (Fox, 2010, p.127-130). This model is a modification of the 2PN IRT model given previously. In the 2PN IRT model, we had an augmented latent variable $Z_{ik} \sim N(a_k\theta_i - b_k, 1)$ where the response $y_{ik} = 1$, if $Z_{ik} > 0$ and $y_{ik} = 0$ otherwise. The probability of examinee $i$ answering item $k$ correctly was given by,

$$P(y_{ik} = 1|\theta_i, \xi_k) = P(Z_{ik} > 0|\theta_i, \xi_k) = \Phi(a_k\theta_i - b_k).$$

The modification to this by Bradlow et al. (1999) is to incorporate the parameter $\gamma_{id(k)}$, which is a person specific testlet effect that is independent of the other ability and item parameters. The testlet parameter $\gamma_{id(k)}$ is considered to be a random effect that models the additional dependence of individuals’ responses to items within testlets. This parameter is incorporated into the 2PN testlet model where the probability of examinee $i$ answering item $k$ correctly is given by,

$$P(y_{ik} = 1|\theta_i, \xi_k, \gamma_{id(k)}) = \Phi(a_k\theta_i - b_k - \gamma_{id(k)}),$$

(5.1)

where $\gamma_{id(k)}$ represents the interaction of item $k$ with person $i$ that is nested in testlet $d(k)$. In this model the augmented latent variable $Z_{ik}$ would be normally distributed with mean $a_k\theta_i - b_k - \gamma_{id(k)}$ and variance 1, where we consider the response $y_{ik} = 1$, if $Z_{ik} > 0$ and $y_{ik} = 0$ otherwise.

The standard interpretations of the parameters $a_k$, $b_k$, and $\theta_i$ still hold as the item discrimination, item difficulty, and examinee ability, respectively. The prior distributions for the ability and item parameters are the same as with the 2PN IRT model. The testlet
parameter $\gamma_{id(k)}$ is taken to have the prior specification,

$$\gamma_{id(k)} \sim N(0, \sigma^2_\gamma),$$  \hspace{1cm} (5.2)

and for each examinee $i$ the $\gamma_{id(k)}$ are taken to be independent across testlets. The testlet variance $\sigma^2_\gamma$, is the same for all testlets.

Before we continue, we should make sure we understand the indices of the testlet parameter $\gamma_{id(k)}$. Each item $k$ belongs to a testlet $d(k) \in \{1, \ldots, L\}$ and for each $i$ the parameter $\gamma_{id(k)}$ is the same for all items in the same testlet. This is how, for a given examinee, the extra dependence of items within the same testlet is modeled, because all items would share the same effect $\gamma_{id(k)}$ in their linear score predictor $m_{ik} = a_k\theta_i - b_k - \gamma_{id(k)}$. That is, for item responses by examinee $i$ to items, say $k$ and $k'$ ($k \neq k'$), that are in the same testlet $d(k) = d(k')$ we would have $m_{ik} = a_k\theta_i - b_k - \gamma_{id(k)}$ and $m_{ik'} = a_k'\theta_i - b_k' - \gamma_{id(k')}$ where $\gamma_{id(k)} = \gamma_{id(k')}$. For each individual $i$ then, there are $L$ testlet parameters. In total the testlet parameters $\gamma$ can be taken as an $(I \times L) \times 1$ vector. Finally, we note that the notation for the testlet parameter $\gamma_{id(k)}$ may vary slightly from source to source. However, this is the notation used by our references Bradlow et al. (1999) and Wainer et al. (2007), and so this is what we shall use.

Wainer et al. (2007) introduce a testlet model based on the 3PL IRT model. This can be given as,

$$P(y_{ik} = 1|\theta_i, \xi_k, c_k, \gamma_{id(k)}) = c_k + (1 - c_k)\Psi(a_k\theta_i - b_k - \gamma_{id(k)}),$$  \hspace{1cm} (5.3)

where $\Psi(.)$ is the standard logistic cdf. With this model, the prior distribution for the testlet parameters is changed to $\gamma_{id(k)} \sim N(0, \sigma^2_{d(k)})$. With this, the variance $\sigma^2_{d(k)}$ of the testlet parameters is different for each testlet. We note that this is different than the two-parameter case when $\sigma^2_\gamma$ is the same across all testlets. As explained in Wainer et al. (2007, p. 136), this is to reflect the possibility that testlets may have different amounts of local
dependence. The probit version giving the 3PN testlet model is given by,

\[ P(Y_{ik} = 1 | \theta_i, \xi_k, c_k, \gamma_{id(k)}) = c_k + (1 - c_k)\Phi(a_k\theta_i - b_k - \gamma_{id(k)}). \] (5.4)

We bother to mention the three parameter testlet model here to aid our discussion of the work done by Sinharay and Johnson (2003) mentioned in the end of section 5.3.2 below. Before returning to this however, we will discuss some of the issues that arise when standard IRT models are fit to data generated under a testlet model.

5.2.2 Issues of Fitting IRT Models to Testlet Data

As mentioned, items within a testlet rely on a common stimulus which yields dependence among an individual’s responses to those items. Standard IRT models assume all item responses of an examinee to be independent conditional on their ability and the item parameters. If say, a 2PN IRT model is fit to data that is generated under a 2PN testlet model then the dependence of item responses within testlets is ignored. Bradlow et al. (1999) discuss the problems with this. They mention that there will be bias in the item discrimination and difficulty parameter estimates, and an overstatement of precision in the ability parameter estimates. The problem with having bias in the item parameter estimates is clear. The issue with the ability parameters is that the spread of the posterior distributions of the abilities under the 2PN IRT model are narrower than would be under the true 2PN testlet model. This causes a problem when computing things like posterior interval estimates. This is because in Bayesian statistics posterior interval estimates are used (like confidence intervals in frequentist statistics) to estimate a range of values a true parameter is likely to be between, and also gauge the accuracy of your point estimate. So if our intervals are too narrow, this can give the researcher a false belief in the accuracy of your estimates.

Bradlow et al. (1999) conduct a simulation to illustrate this point. They generate data under a 2PN testlet model, and both the 2PN testlet model and 2PN IRT model were fitted
to the data using a MCMC Gibbs sampling procedure. They compute posterior sample
means as point estimates of the ability and item parameters under both fitted models. They
compare these two sets of parameter estimates to the true values of the data generating
parameters by computing the mean absolute error (MAE). That is, for say data generating
ability parameters $\theta_1, \ldots, \theta_I$ and posterior estimates $\tilde{\theta}_1, \ldots, \tilde{\theta}_I$, they compute MAE
$\text{MAE} = \sum_{i=1}^I |\tilde{\theta}_i - \theta_i| / I$.

This simulation was carried out for different response data sets generated with different
testlet sizes and values of the testlet variance $\sigma^2_{\gamma}$. It was found that for a given testlet size,
the MAE of the item parameter estimates computed under the 2PN IRT was generally larger
than the item parameter estimates computed under the 2PN testlet model, over the different
values of $\sigma^2_{\gamma}$ used. This shows that, on average, the absolute differences between the item
parameter estimates and true values is greater under the 2PN IRT model, i.e. the 2PN IRT
estimates are farther from the true values than the 2PN testlet estimates. This may not
imply biasedness in the 2PN IRT item parameter estimates right up front, but as we will see
soon it does happen.

Bradlow et al. (1999) also looked at the “overstatement of precision” in the ability
parameter estimates under the 2PN IRT model. Under each model, they computed the
mean posterior interval width (MPIW) over the posterior samples of abilities, where they
computed the mean width of the sample 95% posterior intervals for each $\theta_i$. They also
computed the percentage of true $\theta_i$ values that were contained in the sample 95% posterior
intervals under each model. It was found that the MPIW was generally narrower under the
2PN IRT model. Incidentally, it was also found that the percentage of true $\theta_i$ values actually
contained in the sample 95% posterior intervals was less under the 2PN IRT model than in
the 2PN testlet model.

To illustrate these issues ourselves, we perform a similar simulation as an example. We
consider a test with $I = 500$ examinees and $K = 40$ items. The exam was considered to
have five testlets consisting of eight items each, where items $k = 1, \ldots, 8$ made up the first
testlet, items $k = 9, \ldots, 16$ made up the second testlet, and so on. We generate our response data under a 2PN testlet model given by,

\[
y_{ik} = \begin{cases} 
1, & \text{if } Z_{ik} > 0 \\
0, & \text{if } Z_{ik} \leq 0 
\end{cases} \tag{5.5}
\]

\[Z_{ik} \sim N(a_k \theta_i - b_k - \gamma_{id(k)}, 1)\]

\[\theta \sim N(0, 1),\]

\[a_k \sim TN(0, \infty)(1, .50),\]

\[b_k \sim N(0, 1),\]

\[\gamma_{id(k)} \sim N(0, \sigma^2_\gamma),\]

for $i=1, \ldots, 500$, $k = 1, \ldots, 40$ and $d(k) \in \{1, \ldots, 5\}$.

We generate three different data sets denoted $Y_{\sigma^2_\gamma=0}$, $Y_{\sigma^2_\gamma=1/2}$, and $Y_{\sigma^2_\gamma=1}$ corresponding to three different values of the testlet variance $\sigma^2_\gamma = 0$, $1/2$, and $1$, respectively. The value $\sigma^2_\gamma = 0$ is used as a baseline for comparison since under this the 2PN testlet model is the same as the 2PN IRT model. The values of $\sigma^2_\gamma = 1/2$ and $1$ were picked because these are considered to be plausible values of testlet variance [Wainer et al. 2007]. After generating the three data sets, we fitted the data generating 2PN testlet model to each as well as the 2PN IRT model using MCMC Gibbs sampling. For each fitted model, we draw posterior samples of the ability and item parameters and compute the sample means as posterior parameter estimates.

To examine any bias in the estimates we create difference plots where the difference between the estimates and true values ($\text{estimate} - \text{true value}$) is plotted against the true parameter values. We plot these below in figures 5.1, 5.2, and 5.3. In each figure there are three rows of plots. Each row corresponds to posterior estimates arising under the 2PN testlet model (left side) and 2PN IRT model (right side) fitted to the data generated using each of the three different testlet variances $\sigma^2_\gamma = 0$, $1/2$ and $1$. The top row corresponds to
We start by looking at the item parameter estimates. In figures 5.1 and 5.2 we have the difference plots of the posterior estimates of the discrimination and difficulty parameters, respectively, obtained under fitting each model. In the first row of each figure we see that the plots on the left and right side appear to be very similar. This is to be expected since when $\sigma_\gamma^2 = 0$, then the fitted 2PN testlet and 2PN IRT models are essentially the same. Thus we should expect to get similar posterior estimates under each.

Looking at the second and third rows of figures 5.1 and 5.2 we can see bias in the 2PN IRT posterior estimates of both parameters under the conditions when $\sigma_\gamma^2 = 1/2$ and 1. In particular, looking at figure 5.1 we see that the 2PN IRT discrimination parameter estimates are generally too low, as indicated by many of the points lying below the line $y = 0$. This discrepancy between the true parameter value and estimate is increased as the true discrimination parameter value gets larger. As is mentioned in Bradlow et al. (1999, p. 164) this is to be expected when fitting the 2PN IRT model to data generated under a 2PN testlet model. This is because the 2PN IRT model views “...the variation controlled by $\gamma$ as noise and $a$ is a measure of the correlation of the item with the underlying trait, if the model fits less well (there will be more noise), this correlation will be smaller” (Bradlow et al. 1999, p. 164). In figure 5.2 we see that the 2PN IRT posterior estimates for the difficulty parameters are positively biased for items with difficulty parameter $b < 0$ and negatively biased for items with $b > 0$. We also see that the bias in the 2PN IRT difficulty parameter estimates is increased as $\sigma_\gamma^2$ goes from $1/2$ to 1.

As mentioned earlier, the issue with fitting the 2PN IRT model to testlet data, as it relates to the ability parameters, is the posterior precision. We will address this soon, but we make note that biasedness in the 2PN IRT posterior ability parameter estimates was not mentioned as an issue. Keeping this in mind, if we look at figure 5.3 we see that the left and right plots in all three rows look very similar. Graphically, this seems to indicate that the 2PN IRT ability parameter estimates perform about the same as those under the 2PN
Figure 5.1 Difference plots for the posterior item discrimination parameter estimates computed from fitting the 2PN testlet model (left column) and 2PN IRT model (right column) to data generated from a 2PN testlet model. Three data sets were generated under the 2PN testlet model corresponding to three different values of the testlet variance: $\sigma^2_\gamma = 0, 1/2,$ and 1. The top row corresponds to the data generated using $\sigma^2_\gamma = 0$, the second row to data using $\sigma^2_\gamma = 1/2$, and the last row to data using $\sigma^2_\gamma = 1$. The difference between the estimates and true values ($\text{estimate} - \text{true value}$) is plotted (horizontal) against the true parameter values (vertical).
Figure 5.2 Difference plots for the posterior item difficulty parameter estimates computed from fitting the 2PN testlet model (left column) and 2PN IRT model (right column) to data generated from a 2PN testlet model. Three data sets were generated under the 2PN testlet model corresponding to three different values of the testlet variance: $\sigma^2_\gamma = 0$, $1/2$, and 1. The top row corresponds to the data generated using $\sigma^2_\gamma = 0$, the second row to data using $\sigma^2_\gamma = 1/2$, and the last row to data using $\sigma^2_\gamma = 1$. The difference between the estimates and true values ($\text{estimate} - \text{true value}$) is plotted (horizontal) against the true parameter values (vertical).
As in Bradlow et al. (1999), for each fitted model (2PN testlet and 2PN IRT) and for each data generating value of testlet variance ($\sigma_\gamma^2 = 0, 1/2, \text{ and } 1$) we compute the MAE for the posterior estimates of the ability and item parameters. We put the results in table 5.1. We see that the MAE computed under the 2PN IRT model is greater for both item parameters than the MAE computed under the 2PN testlet model. We also see that the discrepancy increases as the value of $\sigma_\gamma^2$ increases. This quantitatively shows what the difference plots indicated in figures 5.1 and 5.2. For the ability estimates under both models, we see that the MAE values are very close when $\sigma_\gamma^2 = 0$ and $1/2$. When $\sigma_\gamma^2 = 1$ then the MAE was slightly larger under the 2PN IRT model at 0.324 versus 0.282 under the testlet model.

<table>
<thead>
<tr>
<th>$\sigma_\gamma^2$</th>
<th>$\theta$</th>
<th>$a$</th>
<th>$b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Testlet</td>
<td>0.178</td>
<td>0.100</td>
</tr>
<tr>
<td></td>
<td>IRT</td>
<td>0.179</td>
<td>0.103</td>
</tr>
<tr>
<td>1/2</td>
<td>Testlet</td>
<td>0.260</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>IRT</td>
<td>0.266</td>
<td>0.170</td>
</tr>
<tr>
<td>1</td>
<td>Testlet</td>
<td>0.282</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>IRT</td>
<td>0.324</td>
<td>0.265</td>
</tr>
</tbody>
</table>

Table 5.1 Mean absolute errors (i.e. MAE = mean |estimate - true value|) computed between the true parameter values and posterior estimates of the examinee ability and item parameters. The estimates were computed as posterior sample means after fitting the 2PN testlet model and 2PN IRT model to the three data sets generated under the 2PN testlet model with testlet variance being $\sigma_\gamma^2 = 0, 1/2, \text{ and } 1$, respectively. The MAE values for all parameters are very close under both models when $\sigma_\gamma^2 = 0$. When $\sigma_\gamma^2 = 1/2$ and 1, the larger MAE values for the item parameters under the 2PN IRT model shows that, on average, the absolute difference between the parameter estimates and true values is greater than under the (true) 2PN testlet model. The MAE values for the ability parameters under both models are very close when $\sigma_\gamma^2 = 1/2$ and slightly greater under the 2PN IRT model when $\sigma_\gamma^2 = 1$.

We turn now to the issue that was mentioned to arise with the ability parameter estimates when we fit a 2PN IRT model to data generated under a 2PN testlet model. It was said that there is an overstatement of precision in the ability estimates (Bradlow et al. 1999). In this, the spread of the posterior distributions of the abilities under the 2PN IRT model are narrower than would be under the true 2PN testlet model. To show this, under the 2PN
Figure 5.3 Difference plots for the posterior ability parameter estimates computed from fitting the 2PN testlet model (left column) and 2PN IRT model (right column) to data generated from a 2PN testlet model. Three data sets were generated under the 2PN testlet model corresponding to three different values of the testlet variance: $\sigma_{\gamma}^2 = 0$, $1/2$, and $1$. The top row corresponds to the data generated using $\sigma_{\gamma}^2 = 0$, the second row to data using $\sigma_{\gamma}^2 = 1/2$, and the last row to data using $\sigma_{\gamma}^2 = 1$. The difference between the estimates and true values ($\text{estimate} - \text{true value}$) is plotted (horizontal) against the true parameter values (vertical).
testlet and IRT models fitted to the three data sets $Y_{\sigma^2_\gamma=0}$, $Y_{\sigma^2_\gamma=1/2}$, and $Y_{\sigma^2_\gamma=1}$, we compute the mean posterior interval width (MPIW = $\sum_{i=1}^{500}(\hat{F}_{i,0.975} - \hat{F}_{i,0.025})/500$) of the sample 95% posterior intervals as in Bradlow et al. (1999), where $\hat{F}_i$ is the empirical cdf of $\theta_i$ obtained from the corresponding posterior sample. We put the results into Table 5.2. We also compute the percentage of true values $\theta_i$ that were contained in the sample 95% posterior intervals under each of the six simulation conditions. In this, for each fitted model and value of $\sigma^2_\gamma$, we compute $\sum_{i=1}^{500} I(\theta_i \in (\hat{F}_{i,0.025}, \hat{F}_{i,0.975}))/500$ where $I(.)$ is an indicator function. These are given in Table 5.3.

In Table 5.2 we see that the MPIW computed under the 2PN IRT and testlet models is about the same at 0.923 and 0.924, respectively, when $\sigma^2_\gamma = 0$. When $\sigma^2_\gamma = 1/2$ and $\sigma^2_\gamma = 1$ we see the MPIW values under the 2PN IRT model is 0.948 versus 1.211 under the testlet model, and when $\sigma^2_\gamma = 1$ the 2PN IRT MPIW is 0.986 opposed to 1.376. This shows that, on average, the sample 95% posterior intervals computed from fitting the 2PN IRT model are narrower than those under the 2PN testlet model.

We see the effects of this if we look the overall fraction of posterior intervals that cover the true values of $\theta_i$ given in Table 5.3. Since the intervals are constructed to be 95% posterior intervals, then overall we should have around 95% of the true $\theta_i$ values contained in their respective intervals. For all values of $\sigma^2_\gamma$, the percentages of true $\theta_i$ values contained in the 95% posterior intervals computed under the 2PN testlet model are all close to the nominal 95%. Under the 2PN IRT model, when $\sigma^2_\gamma = 0$ the percentage of true $\theta_i$ values contained in the intervals was about 96%, which is close to the nominal percentage and is to be expected. For larger values of $\sigma^2_\gamma$, we see that fewer than the nominal 95% of true values were actually contained in the intervals. In particular, when $\sigma^2_\gamma = 1/2$ there were only about 83% of the true values contained in the posterior intervals, and when $\sigma^2_\gamma = 1$ there was only about 77%.
Table 5.2 Mean Posterior Interval Width (MPIW = $\sum_{i=1}^{500}(\hat{F}_{i,0.975} - \hat{F}_{i,0.025})/500$) of the sample 95% posterior intervals for the examinee abilities computed from posterior samples of the ability parameters drawn from fitting the 2PN IRT model and 2PN testlet model to the three data sets generated under the 2PN testlet model with testlet variance being $\sigma_\gamma^2 = 0$, $1/2$, and 1, respectively. The MPIW values are close under both models when $\sigma_\gamma^2 = 0$, but for $\sigma_\gamma^2 = 1/2$ and 1, the MPIW computed under the 2PN IRT model is smaller. This indicates that, on average, the estimated 95% posterior intervals under the 2PN IRT model are narrower than should be under the true 2PN testlet model.

<table>
<thead>
<tr>
<th>$\sigma_\gamma^2$</th>
<th>2PN IRT</th>
<th>2PN Testlet</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.923</td>
<td>0.924</td>
</tr>
<tr>
<td>1/2</td>
<td>0.948</td>
<td>1.211</td>
</tr>
<tr>
<td>1</td>
<td>0.986</td>
<td>1.376</td>
</tr>
</tbody>
</table>

Table 5.3 Percentage of true values $\theta_i$ that were contained in the sample 95% posterior intervals computed from fitting the 2PN IRT model and 2PN testlet model to the three data sets generated under the 2PN testlet model with testlet variance $\sigma_\gamma^2 = 0$, $1/2$, and 1, respectively. For all values of $\sigma_\gamma^2$, the percentages of true $\theta_i$ values contained in the 95% posterior intervals computed under the 2PN testlet model are all close to the nominal 95%. Under the 2PN IRT model, when $\sigma_\gamma^2 = 0$ the percentage of true $\theta_i$ values contained in the intervals is close to the nominal 95%. When $\sigma_\gamma^2 = 1/2$ and 1, the percentage of $\theta_i$ values contained in the intervals is less than 95%. This is due to the 95% posterior intervals computed under the 2PN IRT model being too narrow.

<table>
<thead>
<tr>
<th>$\sigma_\gamma^2$</th>
<th>2PN IRT</th>
<th>2PN Testlet</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.964</td>
<td>0.960</td>
</tr>
<tr>
<td>1/2</td>
<td>0.828</td>
<td>0.946</td>
</tr>
<tr>
<td>1</td>
<td>0.771</td>
<td>0.942</td>
</tr>
</tbody>
</table>
The discussion in this section, along with these examples, show the problems that arise from fitting a 2PN IRT model to data generated under a 2PN testlet model. These problems being that under a 2PN IRT model, the item parameter estimates will be biased. Also, even though the ability parameter estimates are very close to those obtained from fitting a (true) 2PN testlet model, posterior interval estimates for the abilities will be too narrow. This would mean having an “overconfidence” in the precision of our posterior ability estimates.

Before we address the issue of model checking we will introduce two correlation coefficients that have been used previously as model checking measures for IRT models. We will discuss how these have been used and in particular why they were used. This will be important because we will use one of these as a discrepancy measure under the posterior predictive method to check the fit of the 2PN IRT model to data generated under the 2PN IRT testlet model.

5.3 Introduction to the Biserial and Point-Biserial Correlations

Lord and Novick (1968) give a general discussion of the point biserial correlation and biserial correlation. We give a brief overview of each here. The point biserial correlation, which we denote by $r_{pbis}$, is the correlation between a dichotomous variable and a nondichotomous variable. In IRT models the point biserial correlation of item $k$, denoted $r_{pbis}^{(k)}$, is used to measure the correlation between the binary responses \{$y_{1k}, \ldots, y_{Ik}$\} and the latent traits \{$\theta_1, \ldots, \theta_I$\}. The point biserial correlation can be interpreted much like a standard correlation coefficient. For item $k$, the larger the value of $r_{pbis}^{(k)}$ then the stronger the relationship between the item responses and the latent traits. As we will discuss later $r_{pbis}^{(k)}$ is linked to the discrimination $a_k$, of item $k$.

The formula for $r_{pbis}^{(k)}$ given in Lord and Novick (1968, ch.15) is given in general terms of a dichotomous variable $Y$ and nondichotomous variable $U$. Here we present the formula for $r_{pbis}^{(k)}$ in terms of the dichotomous item response to item $k$ and nondichotomous examinee ability, since we will be using these correlations in an IRT model setting. The formula for
the point biserial correlation is then given as,

\[ r_{pbis}^{(k)} = \frac{\mu^+ - \mu^-}{\sigma_\theta} \sqrt{\pi_k(1 - \pi_k)}, \]  

(5.6)

where \( \mu^+ = E(\theta|Y_k = 1) \), \( \mu^- = E(\theta|Y_k = 0) \), \( \pi_k = E(Y_k) \), and \( \sigma_\theta \) is the standard deviation of \( \theta \). This formula is generally not used in practice, but we give it to motivate the next formula which is used. The sample point biserial correlation, denoted \( r_{pbis}^{(k)} \), is computed by,

\[ r_{pbis}^{(k)} = \frac{M_{p_k} - M_{q_k}}{S_{total}} \sqrt{p_k q_k}, \]  

(5.7)

where \( p_k = \sum_{i=1}^{I} y_{ik}/I \) is the percentage of students correctly answering item \( k \) and \( q_k = 1 - p_k \) is the percentage of students incorrectly answering item \( k \). \( M_{p_k} \) is the mean score for students answering item \( k \) correctly. That is, looking only at the subgroup of examinees that got the item correct, we take the mean of their total scores. So, if we let \( S_i = \sum_{k=1}^{K} y_{ik} \) denote the total test score of examinee \( i \), and \( N_k = \sum_{i=1}^{I} y_{ik} \) denote the total number of examinees that answered item \( k \) correctly, then \( M_{p_k} = (\sum_{i=1}^{I} y_{ik} S_i)/N_k \). Also, \( M_{q_k} = (\sum_{i=1}^{I} (1 - y_{ik}) S_i)/(I - N_k) \) is the mean score for students answering the item \( k \) incorrectly, and \( S_{total} \) is the standard deviation of the examinees’ total scores \( S_i, i = 1, \ldots, I \).

Another correlation coefficient devised by [Pearson (1909)] is the biserial correlation coefficient. This is similar to the point biserial correlation in that it gives the correlation between a dichotomous and nondichotomous variable. However, the difference is that the biserial correlation is used when there is believed to be some continuous latent variable underlying the dichotomy. That is, there is hypothesized to be some latent variable, say \( Z \), where \( Z \) is not observed, but we observe \( Y = 1 \) if \( Z > \delta \) and \( Y = 0 \) if \( Z \leq \delta \) for some value \( \delta \).

Pearson’s biserial correlation is meant to measure the correlation between the original nondichotomous variable and this unobservable latent variable through the observed dichotomous variable \( Y \). An example of this type of setting might be measuring the association between test performance and say, test anxiety. Supposing that anxiety could take any value
on some continuous scale which can not be directly observed, but respondents rate their anxiety as either high or low. Also, suppose we let the dichotomous variable \( Y = 1 \) if test anxiety was high, or \( Y = 0 \) if it was low. Then the correlation between the examinees’ performance and anxiety can be looked at via the dichotomous variable \( Y \) and say, test scores.

Coming back to IRT models, we have such a latent variable that can be incorporated into the model. Recall, we have assumed there to be an augmented latent variable \( Z \) that determines the performance of examinee \( i \) on item \( k \) and the observed item responses are can be thought of as indicator variables. That is, we assume \( Z_{ik} \sim N(a_k \theta_i - b_k, 1) \) where \( Y_{ik} = 1 \) if \( Z_{ik} > 0 \) and \( Y_{ik} = 0 \) if \( Z_{ik} \leq 0 \) for \( i = 1, \ldots, I \) and \( k = 1, \ldots, K \).

The biserial correlation is derived as the product moment correlation (Lord and Novick, 1968, p.339) between the latent continuous variable and the nondichotomous variable, \( Z \) and \( \theta \) in our case, respectively. We should note that Lord and Novick (1968) assume the distribution of the latent variable to be normal, which is the same assumption we are taking with \( Z \). With this, the biserial correlation is given as,

\[
r_{bis} = \frac{\mu^+ - \mu^- \pi(1 - \pi)}{\sigma_\theta \phi(\delta)},
\]

(5.8)

where \( \delta \) is in general the cut off value for the latent variable \( Z \) that defines the dichotomy of \( Y \), and \( \pi = \int_0^\infty \phi(z) \) \( dz \) where \( \phi(\cdot) \) represents the standard normal pdf. The sample biserial correlation coefficient for the \( k^{th} \) item, denoted \( r_{\text{bis}}(k) \), is given as,

\[
r_{\text{bis}}(k) = \frac{M_{p_k} - M_{q_k}}{S_{\text{total}}} \frac{p_k q_k}{\phi(h_k)},
\]

(5.9)

where \( h_k = \Phi^{-1}(p_k) \) with \( \Phi^{-1}(m) \) being the \( m^{th} \) quantile of the standard normal distribution. The quantities \( M_{p_k}, M_{q_k}, p_k, q_k, \) and \( S_{\text{total}} \) are the same as defined for the sample point biserial correlation in (5.7).
5.3.1 Relation between the Biserial and Point Biserial Correlation Coefficients

There is a relation between \( r_{pbis}^{(k)} \) and \( r_{bis}^{(k)} \) (Lord and Novick 1968, p.340) which is given as,

\[
r_{pbis} = r_{bis} \frac{\phi(\delta)}{\sqrt{\pi_k (1 - \pi_k)}}. \tag{5.10}
\]

A parallel equation holds for the sample coefficients \( \hat{r}_{pbis}^{(k)} \) and \( \hat{r}_{bis}^{(k)} \) (Sinharay and Johnson 2003, p.15) and is given as,

\[
\hat{r}_{pbis}^{(k)} = \hat{r}_{bis}^{(k)} \frac{\phi(h_k)}{\sqrt{P_k q_k}}. \tag{5.11}
\]

Both \( \hat{r}_{pbis}^{(k)} \) and \( \hat{r}_{bis}^{(k)}, k = 1, \ldots, K \), are measures of association between the item responses and latent abilities. This implies that these correlations tell us about the discrimination of an item. That is, the higher the correlation (\( r_{pbis}^{(k)} \) or \( \hat{r}_{bis}^{(k)} \)) the better the item discriminates between students with high and low ability. In fact, for the two-parameter normal ogive IRT model it has been shown (Lord and Novick 1968, ch.16) that the biserial correlation can be written directly as a function of the discrimination parameter \( a_k \). This is given as,

\[
r_{bis}^{(k)} = \frac{a_k}{\sqrt{1 + a_k^2}}. \tag{5.12}
\]

Of course, due to the relation in (5.10) this implies \( r_{pbis}^{(k)} \) is directly related to the value of \( a_k \) as well.

5.3.2 Point Biserial Correlation as an IRT Model Checking Measure

Here we discuss how the point biserial correlation can be used as a model checking measure for IRT models. Specifically, we will show how it has been used under a posterior predictive approach to check the fit of a 1PN model to data generated from a 2PN model.
Later we will be investigating the use of the biserial correlation under the posterior predictive method to assess the fit of other IRT models, and this will serve as a good introduction.

Recall from chapter 2 that the 1PN model can be thought of as a special case of the 2PN model where all discrimination parameters are set equal to one. If data were generated from a model where each item has a distinct discrimination parameter value, such as a 2PN model, then this would not be consistent with the assumption of a 1PN model. If one chooses to fit the 1PN model to a set of data, the question of “How appropriate is the assumption of $a \equiv 1$?” could be asked.

Due to the point biserial correlation’s relationship to the discrimination parameters, Albert and Ghosh (2000) used $r_{pbis}$ as a measure to see how well the 1PN model would fit a given data set. Based on (5.10) and (5.12), $r^{(k)}_{pbis}$ can be viewed as a monotonic function of discrimination parameter $a_k$. This is significant because the more variability in the values of $a_k$, $k = 1, \ldots, K$, the more variability there will be in the values of $r^{(k)}_{pbis}$ (and $r^{(k)}_{bis}$) for $k = 1, \ldots, K$.

Albert and Ghosh (2000) used the standard deviation of the sample point biserial correlation coefficients calculated over all items as a measure to check the fit of a 1PN model. Specifically, the 1PN model was fit to a set of data generated under a 2PN model and posterior samples of $\theta$ and $b$ are drawn. From these, data sets $\tilde{Y}$ are simulated from the posterior predictive distribution. For each replicated data set $\tilde{Y}$, the sample point biserial correlation is computed for each item and the standard deviation is computed over the $K$ correlation values, which we denote as $std(\hat{r}_{pbis})$. In effect, this generates a posterior predictive sample of $std(\hat{r}_{pbis})$ values. The value of $std(\hat{r}_{pbis})$ is also computed from the observed data. The observed value is then compared to this posterior predictive sample to determine extremeness under the model.

What we need to remember is that the point biserial (and biserial) correlation is a monotone function of the discrimination parameters and the sample point biserial (and biserial) correlation coefficient is a sample estimate of this. This implies the variability among the
item discrimination parameters would affect the variability among the sample point biserial (and biserial) correlation coefficients. So, if a 1PN model generated the observed data then the discrimination parameter values are all the same and we should expect very little variability among the sample point biserial correlations, i.e. a small observed value of $\text{std}(\hat{r}_{\text{pbis}})$. The posterior predictive distribution of $\text{std}(\hat{r}_{\text{pbis}})$ (under the fitted 1PN model) is used as a reference distribution to compare the observed value. If the observed value of $\text{std}(\hat{r}_{\text{pbis}})$ is large relative to the posterior predictive sample, this implies that the variability of the true discrimination parameters is greater than would be expected for data generated under the 1PN model. Hence, this is taken as evidence that the data was not generated under the 1PN model.

To give an illustration of the example given by Albert and Ghosh (2000), we performed a similar example. We generated a data set $Y_{\text{obs}}$ with $I = 200$ examinees and $K = 40$ items from a 2PN model. We fit the (wrong) 1PN model using an MCMC Gibbs sampling procedure and draw a posterior sample of 1,000 parameters $\{\hat{\theta}^{(r)}, \hat{b}^{(r)}\}_{r=1}^{1000}$. For each $r$, we simulate a data set $\tilde{Y}^{(r)}$ from the posterior predictive distribution as described in chapters 3 and 4. We then compute the corresponding value of $\text{std}(\hat{r}_{\text{pbis}})$, giving us a posterior predictive sample of 1,000 values of $\text{std}(\hat{r}_{\text{pbis}})$ under the fitted 1PN model. We calculate the observed value of $\text{std}(\hat{r}_{\text{pbis}})$ as 0.21. As was done in (Albert and Ghosh, 2000, p.192), we plot a histogram in figure 5.4 which plots the posterior predictive sample of $\text{std}(\hat{r}_{\text{pbis}})$ values as well as the observed value marked by a vertical line.

Visually we can see how much larger the observed value of $\text{std}(\hat{r}_{\text{pbis}})$ is in reference to the posterior predictive sample. More formally, we calculate a posterior predictive $p$-value looking at the proportion of posterior predictive values more extreme than the observed value. That is,

$$P(T(\tilde{Y}) > T(Y_{\text{obs}})) = \frac{1}{1,000} \sum_{r=1}^{1,000} I(T(\tilde{Y}^{(r)}) > T(Y_{\text{obs}})),$$

(5.13)
Figure 5.4 Histogram of a posterior predictive sample of the standard deviations of sample point biserial correlations ($std(\hat{r}_{pbis})$) calculated over all items on 1,000 replicated posterior predictive data sets. The observed data was generated under a 2PN IRT model and the 1PN IRT model was fitted. The observed value of $std(\hat{r}_{pbis})$ calculated from the observed data is marked by the vertical line. We see that the vertical line sits far to the right of the histogram, indicating the observed value of $std(\hat{r}_{pbis})$ is very large in reference to the posterior predictive sample under the 1PN IRT model. Visually, this indicates that the 1PN IRT model is not a good fit to the observed data.

where $T(.) = std(\hat{r}_{pbis})$ and $I(.)$ is an indicator function. The posterior predictive $p$-value is 0, representing essentially no chance of observing a standard deviation value of $\hat{r}_{pbis}$ as extreme as 0.21 under the 1PN model. This suggests that the 1PN model is not suitable for describing the data.

Toribio (2006) conducted a simulation study to investigate the power of this procedure. In this, 100 data sets were simulated under the 2PN model and the 1PN model was fitted to each. A posterior predictive $p$-value was computed for each data set in a similar manner as we did in our example above. The results showed that all $p$-values were significant at the 0.05 threshold level with nearly all being 0.

This shows that the standard deviation of the point biserial correlation is an effective measure in determining the fit of a 1PN model under the posterior predictive method. The biserial correlation, which is proportional to the point biserial correlation from (5.10), has also been investigated as a measure to determine the fit of IRT models. Sinharay and
Johnson (2003) conducted simulation studies using the biserial correlation to check the fit of a 1PL model to data generated under the 2PL and 3PL models. In particular, they used the standard deviation of the sample biserial correlation in a posterior predictive approach similar to what was done by Albert and Ghosh (2000). The standard deviation of the sample biserial correlation was found to be a very effective measure in detecting the lack of fit of the 1PL model when fit to data generated under the 2PL, or 3PL models.

5.4 Checking Model fit of IRT Models to Testlet Data

Along with the studies of Sinharay and Johnson (2003) just mentioned, they also conducted a simulation study to check the ability of the sample biserial correlations to detect the inadequacy of the 3PL model when fit to data generated under the 3PL testlet model. In particular, they generated response data from a 3PL testlet model similar to what we defined in (5.3), in which there were responses of \( I = 2,500 \) examinees to \( K = 35 \) items divided into 6 testlets of 5 items each. Items 1 through 5 made up the first testlet, items 6 through 10 made up the second testlet, and so on. They simulated the data such that each of the six testlets had a separate testlet standard deviation, \( \sqrt{\sigma^2_{d(k)}} \), which are 0, 0.3, 0.5, 0.7, 0.85, and 1.

One-hundred separate data sets were generated under the 3PL testlet model and the 3PL IRT model was fit to each using an MCMC procedure. For each iteration, a posterior predictive sample of the standard deviation, maximum, and mean of the sample biserial correlations was computed and compared to the corresponding observed values computed on the data. The specific simulation results were not given, but it is stated that “...the biserial correlations fail to detect any inadequacy of the 3PL” (Sinharay and Johnson 2003). It is also stated in their conclusion that the biserial correlations have “no power” with regards to this.

Being motivated by the work of Sinharay and Johnson (2003), we are interested in investigating the usefulness of some of the model checking procedures we have used so far in
assessing the fit of IRT models fit to data generated under a testlet model. In particular, we are interested in applying the posterior predictive and PPPS model checking procedures. We will look at assessing the fit of a 2PN IRT model to data generated under the 2PN testlet model. We choose these models because the 2PN setting is a simpler set of models (for both IRT and testlet) to work with to start this type of initial investigation.

The posterior predictive method is widely used in Bayesian IRT model checking due to its ease of implementation so it seems reasonable to want to use this method. We also want to investigate the approach taken by Sinharay and Johnson (2003), in a two-parameter setting. The PPPS method performed well in chapter 4 in person fit checking, so we are interested in seeing if it can be effective in detecting lack of fit in this case. In employing each method though, we first need to determine what discrepancy measures to use. We will discuss which measures we will use under each method and provide some justification for why we chose them. We then finish the chapter by presenting the results of two simulation studies conducted to investigate each method.

5.4.1 Discrepancy Measures Used Under the PP Method

Based on the discussion and example in section 5.3.2, a fact that we have is that the values of the sample biserial and point biserial correlations $\hat{r}_{\text{bis}}^{(k)}$ and $\hat{r}_{\text{pbis}}^{(k)}$ are influenced by the data generating values of the item discrimination parameters. Under the posterior predictive method we generate data sets $\tilde{Y}$ from the posterior predictive distribution under the fitted model by simulating posterior samples of parameters and using these to generate the data values. In this, the “data generating” values of the item discrimination parameters are the posterior values simulated under the fitted model.

Now, recall from section 5.2.2 that when the 2PN IRT model was fit to the 2PN testlet data $Y_{\sigma^2=1/2}$, and $Y_{\sigma^2=1}$ there was noticeable negative bias in the corresponding item discrimination parameter estimates (e.g. see figure 5.1). Due to this discrepancy between the 2PN IRT posterior values of $a_k$ and the true values, and what was just mentioned above, $\hat{r}_{\text{bis}}^{(k)}$
and \( \hat{r}_{bis}^{(k)} \) and \( \hat{r}_{pbis}^{(k)} \) would seem logical choices of discrepancy measures to use under the PP method. If there was a discrepancy between the simulated posterior values of \( a_k \) and the true values, it stands to reason that there would be a discrepancy between the values of \( \hat{r}_{bis}^{(k)} \) and \( \hat{r}_{pbis}^{(k)} \) computed on the posterior predictive data \( \hat{Y} \) and the observed data.

We will choose to use the sample biserial correlation \( \hat{r}_{bis}^{(k)} \). This is because the two measures are proportional to one another, as per (5.11). Also, from (5.12) \( r_{bis}^{(k)} \) can be conveniently computed as a function of \( a_k \) under the 2PN IRT model. We will make use of this latter part in the next section when we discuss how we will use the PPPS method.

As mentioned at the beginning of this section, Sinharay and Johnson (2003) conducted a simulation study to check the ability of the sample biserial correlations to detect the inadequacy of the 3PL model when fit to data generated under the 3PL testlet model. Specifically, they looked at the standard deviation of the sample biserial correlations as well as their mean and maximum as discrepancy measures, which we denote as \( sd(\hat{r}_{bis}) \), \( max(\hat{r}_{bis}) \), and \( mean(\hat{r}_{bis}) \), respectively. These measures were found not to be effective in detecting the lack of fit of the 3PL IRT model under the PP method. We will investigate the effectiveness of these measures in detecting the lack of fit of the 2PN IRT model to 2PN testlet data. In a moment we will look at an example to give an initial idea of how they will perform, but first we will give a little more justification as to why we think these measures might be useful.

In figure 5.1 there was negative bias in the 2PN IRT item discrimination parameter estimates as indicated by the majority of the points lying below the line \( y = 0 \). This negative bias indicates the spread of the 2PN IRT item discrimination estimates is less than the spread of the true values of \( a_k \), \( k = 1, \ldots, K \). We can see this more clearly in figure 5.5 where we plot scatter plots of the true values of \( a_k \) versus the 2PN IRT posterior estimates after fitting the 2PN IRT model to the data sets \( Y_{\sigma^2_\gamma=0} \), \( Y_{\sigma^2_\gamma=1/2} \), and \( Y_{\sigma^2_\gamma=1} \). The left plot corresponds to the posterior estimates obtained from fitting the 2PN IRT model to \( Y_{\sigma^2_\gamma=0} \), the middle plot corresponds to \( Y_{\sigma^2_\gamma=1/2} \), and the right plot to \( Y_{\sigma^2_\gamma=1} \). The line \( y = x \) is drawn for reference.
Figure 5.5 Scatter plots of true values against posterior estimates of the item discrimination parameters $a_k$, obtained from fitting a 2PN IRT model to the three data sets $Y_{\sigma^2=0}$, $Y_{\sigma^2=1/2}$, and $Y_{\sigma^2=1}$, generated under the 2PN testlet model. The left plot corresponds to the posterior estimates obtained from fitting the 2PN IRT model to $Y_{\sigma^2=0}$, the middle plot corresponds to $Y_{\sigma^2=1/2}$, and the right plot to $Y_{\sigma^2=1}$. The line $y = x$ is drawn for reference. Similar to figure 5.1, the negative bias in the 2PN IRT posterior estimates of $a_k$ when $\sigma^2 = 1/2$ and $\sigma^2 = 1$ is evident from the majority of the points lying below the line $y = x$. Looking at the horizontal (true value) and vertical (posterior estimate) scales, we see that the spread of the posterior estimates when $\sigma^2 \neq 0$ is less than the true values. Since the sample values of $r_{bis}^{(k)}$ are influenced by the “data generating” values of $a_k$, summary measures such as the standard deviation, maximum, and mean of the $r_{bis}^{(k)}$ values may be useful under the PP method in indicating a lack of fit with the 2PN IRT model.

The negative bias of the estimates is indicated by the majority of the points lying below the line $y = x$, in the middle and right plots. If we look at the vertical axes we see that the spread of the true values of $a_k$ go from about 0 to 2.6. If we look at the horizontal axis in the middle and right plots we see that the overall spread of the 2PN IRT estimates is less than the true values, being from about 0 to 2.3 and 0 to 2.0, respectively. Now, the estimates are just the means of the posterior samples $\bar{a}_k = (\bar{a}_k^{(1)}, \ldots, \bar{a}_k^{(1000)})$ of each discrimination parameter. However, what this implies is that if the “center” of the posterior sample of $a_k$ is less than the corresponding true value, then the majority of the other sample values will be also. So, we might expect that for a given vector of posterior discrimination sample values $(\bar{a}_1^{(l)}, \ldots, \bar{a}_K^{(l)})$, $l = 1, \ldots, 1000$, the spread of these values will be less than that of the true values of $a_k$, $k = 1, \ldots, K$.

So, if the spread of the posterior values of $a_k$, $k = 1, \ldots, K$ under the fitted 2PN IRT
model is different than the true values, then we can expect the spread of the values of $\hat{r}_{bis}^{(k)}$ computed on the posterior predictive data $\tilde{Y}$ to be different than the spread of $\hat{r}_{bis}^{(k)}$ computed on the observed data. Due to this, we might expect that the standard deviation might be useful in detecting lack of fit. Similarly, in figures 5.1 and 5.5 we see the greatest discrepancy between the maximum discrimination parameter estimate and true value. So, the maximum of the $\hat{r}_{bis}^{(k)}$ values seems natural. Also, if the majority of the posterior discrimination parameters are less than the true values then we might expect to see a discrepancy between the means of the posterior predictive and observed $\hat{r}_{bis}^{(k)}$ values.

Now we will look at an example to see the discrepancy between each of these measures computed on data generated under the true 2PN testlet model, and the posterior predictive distribution computed from fitting the 2PN IRT model. To do this we use the data $Y_{\sigma^2=1}$ that was generated under the 2PN testlet model in our simulation example in section 5.2.2 and compute "observed" values of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$. Using the samples of posterior parameters that were generated from fitting the 2PN IRT model to $Y_{\sigma^2=1}$, we compute a sample of 1,000 posterior predictive response data sets $\tilde{Y}(l)$, $l = 1, \ldots, 1000$. Over each of these data sets we compute the three measures, giving us a posterior predictive sample of size 1,000 for each measure. We use data generated with the value of $\sigma^2 = 1$, because under this we saw the greatest discrepancy between the posterior parameter values under both models. So, if the measures can not indicate a lack of fit here, then they will not likely be able to for smaller data generating values of $\sigma^2$.

In figure 5.6, from left to right, we plot histograms of these posterior predictive samples of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$. We plot vertical lines at the observed value of the measures computed from the data $Y_{\sigma^2=1}$. In the left and middle plots we see the observed values of $sd(\hat{r}_{bis})$ and $max(\hat{r}_{bis})$ are in the left tails of the histograms of the posterior predictive samples. The right plot shows the observed value of $mean(\hat{r}_{bis})$ in the center of the posterior predictive sample. We calculate the $p$-values as in (5.13) and get 0.905, 0.908, and 0.470 for $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$, respectively. This indicates
that there is some promise in $sd(\hat{r}_{bis})$ and $max(\hat{r}_{bis})$ in being effective discrepancy measures under the posterior predictive method, but $mean(\hat{r}_{bis})$ may not be very effective.

Figure 5.6 From left to right, histograms of posterior predictive samples of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$ computed after fitting a 2PN IRT model to data $Y_{\sigma^2=1}$ generated under a 2PN testlet model. The vertical lines indicate the “observed” values of the measures computed on the observed data $Y_{\sigma^2=1}$. The observed values of $sd(\hat{r}_{bis})$ and $max(\hat{r}_{bis})$ lie in the left tails of their respective posterior predictive distributions. This discrepancy between the observed and posterior predictive values of $sd(\hat{r}_{bis})$ and $max(\hat{r}_{bis})$ implies these measures may be useful in indicating a lack of fit with the 2PN IRT model. The observed value of $mean(\hat{r}_{bis})$ lies in the center of the posterior predictive distribution, implying this measures may be ineffective in detecting lack of fit.

5.4.2 Discrepancy Measures Used Under the PPPS Method

Under the posterior predictive method, as has been discussed in chapter 3 and 4, the discrepancy measures used need to be functions involving the data due to the use of the posterior predictive distribution. The PPPS method allows for discrepancy measures that are not direct functions of the data. That is, we can use measures that are functions of parameters only if we wish. Note that our motivation for using the $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$ measures under the posterior predictive method was due to the bias that arose in the posterior discrimination parameters when fitting a 2PN IRT model. With this in mind, under the PPPS method we can look at discrepancy measures that are functions of the item discrimination parameters directly.

Before continuing, lets recall the general procedure for model checking we take under the
PPPS method. For a chosen discrepancy measure $T$ and a set of observed data $Y_{\text{obs}}$, we first compute a posterior sample of $T$ given $Y_{\text{obs}}$ under the fitted model. Then we simulate say $N$ data sets $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(N)}$, under the prior predictive distribution of the model and compute a posterior sample of $T$ given each data set $Y_{\text{rep}}^{(n)}$. The posterior sample of $T$ computed given $Y_{\text{obs}}$ is then compared to the $N$ posterior samples of $T$ computed given the replicated data sets $Y_{\text{rep}}^{(n)}$, $n = 1, \ldots, N$. The main idea being that if the posterior sample of $T$ computed given $Y_{\text{obs}}$ is significantly different than the posterior samples of $T$ computed given $Y_{\text{rep}}^{(n)}$, this is evidence that the data $Y_{\text{obs}}$ was not generated under the model. An effective discrepancy measure $T$ should then be one that if data $Y$ was not generated under the model, the computed posterior sample of $T$ given $Y$ is significantly different than those we get from data generated under the model.

Our method of comparing the posterior sample of $T$ arising under $Y_{\text{obs}}$ and the set of posterior samples arising under $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(N)}$ was to look at the 5 sample quantiles, 0.05, 0.25, 0.50, 0.75, and 0.95. Specifically, after simulating the posterior samples of the measure based on the replicated prior predictive data sets $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(N)}$, we computed the 5 aforementioned quantiles over each of these samples. We then computed the mean quantile vector of the five quantiles over these posterior samples. Then we compute the Euclidean distances between the five quantiles of all the posterior samples and this mean quantile vector. The Euclidean distance $e_0$, computed from the posterior sample obtained given $Y_{\text{obs}}$, is then compared to the other Euclidean distances of the posterior samples. If $e_0$ is significantly large relative to the other Euclidean distances, then this implies the posterior sample of $T$ given $Y_{\text{obs}}$ is significantly different from the posterior samples of $T$ given $Y_{\text{rep}}^{(n)}$, $n = 1, \ldots, N$. Hence, there is an indication of model lack of fit with $Y_{\text{obs}}$.

So far we have seen that the posterior discrimination parameters simulated from fitting the 2PN IRT model to 2PN testlet data $Y$ tend to be negatively biased. A question for us to ask would be, “How would the posterior discrimination parameters simulated from fitting the 2PN IRT model to $Y$ compare to those simulated when the 2PN IRT model is
fit to $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(N)}$? As discussed in the previous section 5.4.1, the negative bias in the 2PN IRT posterior discrimination parameters implies that for of a given vector of posterior discrimination sample values $(\tilde{a}_1^{(l)}, \ldots, \tilde{a}_K^{(l)})$, $l = 1, \ldots, 1000$, the spread of these values will be less than that of the true values of $a_k$, $k = 1, \ldots, K$. Similarly, if the posterior discrimination parameters arising from fitting the 2PN IRT model to $Y$ are generally smaller than those obtained from fitting the 2PN IRT model to $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(N)}$, then it would make sense that the variability among those parameters would be smaller. This motivates using the standard deviation of the discrimination parameters, $sd(a) = \sqrt{\sum_{k=1}^{K} (a_k - \bar{a})^2 / (K - 1)}$, as a discrepancy measure, just as it is used with the sample biserial correlations.

We will investigate this shortly, but first we will consider using two other discrepancy measures computed on the posterior discrimination parameters. Since the bias in the posterior discrimination parameters simulated from fitting the 2PN IRT model to 2PN testlet data increases as the item discrimination gets larger, another seemingly logical choice of discrepancy measure would be the maximum, $max(a) = a_{(K)}$. We will also look at the mean of the $a_k$, $mean(a) = \sum_{k=1}^{K} a_k / K$, since it might be reasonable to expect that due to the smaller values, the means of the posterior discrimination parameters will be smaller.

Before going on, we first clarify how we will compute these measures. Using our MCMC procedure, we will draw a posterior sample of the discrimination parameters of length say 1,000, which will give us a $K \times 1000$ matrix where each column gives us a vector $\tilde{a}^{(l)} = (\tilde{a}_1^{(l)}, \ldots, \tilde{a}_K^{(l)})^t$, $l = 1, \ldots, 1000$. We will compute the standard deviation, maximum, and mean over each column which will in effect give us a posterior sample of length 1,000 of each measure.

We start by initially investigating whether the standard deviation, maximum, and mean of the posterior discrimination parameters will be effective discrepancy measures under the PPPS method. We perform a simulation where we generate an observed $500 \times 40$ response matrix $Y_{obs}$ under the 2PN testlet model described in (5.5) with $\sigma^2_\gamma = 1$, and fit the 2PN IRT model. Using an MCMC procedure we draw posterior samples of the discrimination
parameters under the fitted model and compute the posterior samples of the three measures. Again, if the posterior discrimination parameters simulated from fitting the 2PN IRT model to 2PN testlet data tend to be smaller, we would expect to see this reflected in the posterior samples of the measures. To check this, we also simulate 10 prior predictive response matrices, \( Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)} \), under the 2PN IRT model and fit the 2PN IRT model to each. For each, we then compute posterior samples of the three discrepancy measures as done for the observed data \( Y_{obs} \).

In figures 5.7, 5.9, and 5.11, we summarize the posterior samples of the measures by plotting box plots of the 90% posterior intervals. The box plots extend from the 5th to 95th quantile and the inner box indicates the 25th, 50th and 75th quantiles. The first box plot, labeled Testlet, is from the posterior sample of the measure obtained by fitting the 2PN IRT model to the observed testlet data \( Y_{obs} \). The next 10 box plots (numbered 1 to 10) are from the posterior samples obtained by fitting the 2PN IRT model to the 10 2PN IRT prior predictive response data sets \( Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)} \). The last box plot, labeled Mean, plots the mean vector of the five quantiles (0.5, 0.25, 0.50, 0.75, 0.95) indicated in the previous 10 box plots.

Recall, one last time, that in the PPPS method we judge model lack of fit to a set of observed data \( Y \) by seeing how different the posterior sample of the discrepancy measure given \( Y \), is to posterior samples of the measure given data replicated under the model. The “difference” is measured by seeing how close the 5 quantiles of the posterior sample given \( Y \) are to the mean quantiles, in relation to those of the posterior samples given the replicated data. The box plots in figures 5.7, 5.9, and 5.11 allow us to do this visually.

If we want to quantify this difference between the posterior samples, we compute the Euclidean distances. In figures 5.8, 5.10, and 5.12 we plot the Euclidean distances computed between the 5 quantiles of each sample and the mean quantile vector. The first 10 points in all 3 figures, numbered 1 through 10, indicate the Euclidean distances corresponding to each of the posterior samples of the measure obtained given \( Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)} \). The last
star shaped point labeled Testlet, indicates the Euclidean distance $e_0$ computed using the posterior sample simulated given the observed testlet data $Y_{obs}$.

Looking at figures 5.7, 5.9, and 5.11 we see that for each measure the Testlet box plots of the posterior samples corresponding to the testlet data $Y_{obs}$ are shifted down from those corresponding to $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$. This shows that the standard deviation, maximum, and mean of the posterior discrimination parameters given $Y_{obs}$ are overall smaller than those arising given $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$. Also, we note that the Testlet box plots are shifted farther away from the Mean box plots than the middle 10 box plots. So, at least visually, under the PPPS method this would indicate that according to the discrepancy measures $sd(a)$, $max(a)$, and $mean(a)$, the data $Y_{obs}$ is not likely to have been generated under the 2PN IRT model. In figures 5.8, 5.10, and 5.12 we in fact see that the largest computed Euclidean distances correspond to the posterior samples of the measures given $Y_{obs}$.

This implies that the measures $sd(a)$, $max(a)$, and $mean(a)$ may perform well in detecting the lack of fit of the 2PN IRT model when using this PPPS method. Of course this example used only 10 posteriors simulated given prior predictive data, but was just to give us an initial idea of how effective these measures might be. In the next section we will perform a simulation study to investigate both the type I error rate and detection rate of the measures, using 100 prior predictive data sets.

5.5 Simulation Study

In this section we will discuss the results of a simulation study that was conducted to investigate the effectiveness of the aforementioned sets of discrepancy measures under the PP and PPPS methods for assessing the fit of the 2PN IRT model to 2PN testlet data. Under the PP method, we will use the measures discussed in section 5.4.1. Those were the standard deviation, maximum, and mean of the sample biserial correlations; $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$, respectively. Under the PPPS method we will use the measures discussed in section 5.4.2. Those were the standard deviation, maximum, and mean of the
Figure 5.7 Box plots giving the 90% posterior intervals of the samples of the computed standard deviation on the discrimination parameters, denoted as $sd(a)$. The box plots extend from the 5th to 95th quantile with the inner boxes indicating the 25th, 50th and 75th quantiles. The first box plot, labeled Testlet is from the posterior sample obtained by fitting the 2PN IRT model to the observed testlet data $Y_{obs}$. The next 10 box plots, labeled 1 to 10, are from posterior samples obtained by fitting the 2PN IRT model to 10 prior predictive data sets $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$ generated under the 2PN IRT model. The last box plot, labeled Mean, plots the mean vector of the five quantiles of the 10 previous box plots. These box plots allow us to visually compare the 5 quantiles (0.05, 0.25, 0.50, 0.75, and 0.95) of the the posterior sample of $sd(a)$ given $Y_{obs}$ to the mean quantiles, in relation to those of the posterior samples given $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$. We see from the Testlet box plot that the 5 quantiles of the posterior sample of $sd(a)$ given $Y_{obs}$ are, respectively, farther away from the mean quantiles than those corresponding to $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$. The discrepancy between the posterior samples of $sd(a)$ given $Y_{obs}$ and $Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)}$ indicates that may be an effective discrepancy measure in detecting the lack of fit of the 2PN IRT model under the PPPS method.
Figure 5.8 We plot the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of figure 5.7) of each posterior sample of the measure \( sd(a) \), and the mean quantile vector. The first 10 points, labeled 1 to 10, are the Euclidean distances between the 5 quantiles of the 10 prior predictive posterior samples of \( sd(a) \) under the 2PN IRT model, and the mean quantile vector over these 10 samples. The last point, labeled *Testlet*, is the Euclidean distance between the 5 quantiles of the posterior sample under the fitted 2PN IRT model to the observed testlet data \( Y_{obs} \), and the mean quantile vector. In figure 5.7 we saw that “…the 5 quantiles of the posterior sample of \( sd(a) \) given \( Y_{obs} \) are, respectively, farther away from the mean quantiles than those corresponding to \( Y_{rep}^{(1)}, \ldots, Y_{rep}^{(10)} \).” Here, this is confirmed since the largest Euclidean distance is indicated by the last point which corresponds to the posterior sample of \( sd(a) \) given \( Y_{obs} \).
Figure 5.9 Box plots giving the 90% posterior intervals of the samples of the computed maximum on the discrimination parameters, denoted as \(\text{max}(a)\). The box plots extend from the 5th to 95th quantile with the inner boxes indicating the 25th, 50th and 75th quantiles. The first box plot, labeled \textit{Testlet} is from the posterior sample obtained by fitting the 2PN IRT model to the observed testlet data \(Y_{\text{obs}}\). The next 10 box plots, labeled 1 to 10, are from posterior samples obtained by fitting the 2PN IRT model to 10 prior predictive data sets \(Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}\) generated under the 2PN IRT model. The last box plot, labeled \textit{Mean}, plots the mean vector of the five quantiles of the 10 previous box plots. These box plots allow us to visually compare the 5 quantiles (0.05, 0.25, 0.50, 0.75, and 0.95) of the the posterior sample of \(\text{max}(a)\) given \(Y_{\text{obs}}\) to the mean quantiles, in relation to those of the posterior samples given \(Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}\). We see from the \textit{Testlet} box plot that the 5 quantiles of the posterior sample of \(\text{max}(a)\) given \(Y_{\text{obs}}\) are, respectively, farther away from the mean quantiles than those corresponding to \(Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}\). The discrepancy between the posterior samples of \(\text{max}(a)\) given \(Y_{\text{obs}}\) and \(Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}\) indicates that may be an effective discrepancy measure in detecting the lack of fit of the 2PN IRT model under the PPPS method.
Figure 5.10 We plot the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of figure 5.9) of each posterior sample of the measure $\max(a)$, and the mean quantile vector. The first 10 points, labeled 1 to 10, are the Euclidean distances between the 5 quantiles of the 10 prior predictive posterior samples of $\max(a)$ under the 2PN IRT model, and the mean quantile vector over these 10 samples. The last point, labeled Testlet, is the Euclidean distance between the 5 quantiles of the posterior sample under the fitted 2PN IRT model to the observed testlet data $Y_{\text{obs}}$, and the mean quantile vector. In figure 5.9 we saw that “... the 5 quantiles of the posterior sample of $\max(a)$ given $Y_{\text{obs}}$ are, respectively, farther away from the mean quantiles than those corresponding to $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$.” Here, this is confirmed since the largest Euclidean distance is indicated by the last point which corresponds to the posterior sample of $\max(a)$ given $Y_{\text{obs}}$. 
Figure 5.11 Box plots giving the 90% posterior intervals of the samples of the computed mean on the discrimination parameters, denoted as $\text{mean}(a)$. The box plots extend from the 5th to 95th quantile with the inner boxes indicating the 25th, 50th and 75th quantiles. The first box plot, labeled Testlet is from the posterior sample obtained by fitting the 2PN IRT model to the observed testlet data $Y_{\text{obs}}$. The next 10 box plots, labeled 1 to 10, are from posterior samples obtained by fitting the 2PN IRT model to 10 prior predictive data sets $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$ generated under the 2PN IRT model. The last box plot, labeled Mean, plots the mean vector of the five quantiles of the 10 previous box plots. These box plots allow us to visually compare the 5 quantiles (0.05, 0.25, 0.50, 0.75, and 0.95) of the the posterior sample of $\text{mean}(a)$ given $Y_{\text{obs}}$ to the mean quantiles, in relation to those of the posterior samples given $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$. We see from the Testlet box plot that the 5 quantiles of the posterior sample of $\text{mean}(a)$ given $Y_{\text{obs}}$ are, respectively, farther away from the mean quantiles than those corresponding to $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$. The discrepancy between the posterior samples of $\text{mean}(a)$ given $Y_{\text{obs}}$ and $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$ indicates that may be an effective discrepancy measure in detecting the lack of fit of the 2PN IRT model under the PPPS method.
Figure 5.12 We plot the Euclidean distances between the 5 quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 (indicated in the box plots of figure 5.11) of each posterior sample of the measure $\text{mean}(a)$, and the mean quantile vector. The first 10 points, labeled 1 to 10, are the Euclidean distances between the 5 quantiles of the 10 prior predictive posterior samples of $\text{mean}(a)$ under the 2PN IRT model, and the mean quantile vector over these 10 samples. The last point, labeled Testlet, is the Euclidean distance between the 5 quantiles of the posterior sample under the fitted 2PN IRT model to the observed testlet data $Y_{\text{obs}}$, and the mean quantile vector. In figure 5.11 we saw that “...the 5 quantiles of the posterior sample of $\text{mean}(a)$ given $Y_{\text{obs}}$ are, respectively, farther away from the mean quantiles than those corresponding to $Y_{\text{rep}}^{(1)}, \ldots, Y_{\text{rep}}^{(10)}$.” Here, this is confirmed since the largest Euclidean distance is indicated by the last point which corresponds to the posterior sample of $\text{mean}(a)$ given $Y_{\text{obs}}$. 
item discrimination parameters; \( sd(a) \), \( max(a) \), and \( mean(a) \), respectively. We will first explain the simulation conditions under which the data was generated. Then we will discuss the separate results that were obtained under both methods.

Under both methods we will simulate the “observed” response data sets \( Y_{obs} \) under the 2PN testlet model given in (5.5). These will consist of responses of \( I = 500 \) examinees to \( K = 40 \) items. The items will be divided into five testlets consisting of eight items each, where items \( k = 1, \ldots, 8 \) make up the first testlet, items \( k = 9, \ldots, 16 \) make up the second testlet, and so on.

As discussed in section 5.2.2, the discrepancy between the posterior parameter distributions of the 2PN IRT and 2PN testlet models fitted to 2PN testlet data \( Y \), is influenced by the data generating value of the testlet variance \( \sigma^2_{\gamma} \). We will consider four simulation conditions in which response data \( Y_{obs} \) will be simulated with testlet variance \( \sigma^2_{\gamma} = 0 \), 0.50, 0.75, and 1. Even though when the testlet variance \( \sigma^2_{\gamma} = 0 \) the 2PN testlet model is the same as the 2PN IRT model, we include this as a baseline for comparison purposes.

Under each model checking method, we investigated the type I error rate and detection rate. In studying the type I error rate, we are examining how often the measures under each method will wrongly indicate a lack of fit of the data generating model to \( Y_{obs} \). In studying the detection rates, we are investigating how often the measures will correctly indicate a lack of fit of the 2PN IRT model to \( Y_{obs} \) under both methods. Here we start with discussing the results under the posterior predictive method followed by those under the PPPS method.

## 5.5.1 Posterior Predictive Simulation

Here we will discuss the results of the simulation study employing the posterior predictive method in which we use the measures \( sd(\hat{r}_{bis}) \), \( max(\hat{r}_{bis}) \), and \( mean(\hat{r}_{bis}) \) discussed in section 5.4.1. The first part of this study was to investigate the type I error rates of the measures. The second part of the study was to look at the detection rates when a 2PN IRT model was fit to data generated under a 2PN testlet model.
5.5.1.1 Type I Error Rate Under the Posterior Predictive Method

We start by discussing the process we used in investigating the type I error rate of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$ under the PP method. For each given value of testlet variance $\sigma^2_\gamma = 0, 0.50, 0.75, \text{and } 1$, we simulated a response data set $Y_{obs}$ under the 2PN testlet model in (5.5) and computed the observed value of each measure. Then we fit the data generating model using an MCMC procedure and drew posterior samples of 1,000 item and ability parameters. These were then used to compute a sample of 1,000 posterior predictive response data sets, $\tilde{Y}(l)$ for $l = 1, \ldots, 1000$. The three measures were computed for each of the posterior predictive data sets. That is, on each data set $\tilde{Y}(l)$ we first computed the sample biserial correlations $\hat{r}_{bis}^{(k)}$, $k = 1, \ldots, 40$, and then computed the standard deviation, maximum, and mean of these. Thus, giving posterior predictive samples of length 1,000 for each measure.

We then used the observed values of the measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$ and their respective posterior predictive samples to compute posterior predictive $p$-values,

$$P(T(\tilde{Y}) > T(Y_{obs})) = \frac{1}{1,000} \sum_{r=1}^{1,000} I(T(\tilde{Y}(l)) > T(Y_{obs})),$$

(5.14)

where $T(.) = sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, or $mean(\hat{r}_{bis})$ and $I(.)$ is an indicator function. Posterior predictive $p$-values greater than 0.95 were considered to indicate inadequacy of the fitted model to the data.

This procedure was iterated 100 times under each of the four values of testlet variance considered. Table 5.4 displays the total number of times out of 100 simulations the PP $p$-values were greater than 0.95, indicating a lack of fit of the data generating model with the observed data (i.e. the type I error rate). Each row gives the rates for the three measures corresponding to a particular value of $\sigma^2_\gamma$ used in the simulation.
Type I Error Rate Measures

<table>
<thead>
<tr>
<th>$\sigma_\gamma^2$</th>
<th>$sd(\hat{r}_{bis})$</th>
<th>$max(\hat{r}_{bis})$</th>
<th>$mean(\hat{r}_{bis})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.50</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.75</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 5.4 Type I error rates, or number of times out of 100 simulations the 3 discrepancy measures $sd(\hat{r}_{bis})$, $max(\hat{r}_{bis})$, and $mean(\hat{r}_{bis})$, evaluated using the PP method, indicated a lack of fit with a $p$-value $> 0.95$. Response data $Y_{obs}$ was generated from a 2PN testlet model under 4 simulation conditions with testlet variance $\sigma_\gamma^2 = 0$, 0.50, 0.75, and 1, and the data generating model was fitted.

We see that all type I error rates are very small. The $sd(\hat{r}_{bis})$ and $mean(\hat{r}_{bis})$ discrepancy measures did not indicate any lack of fit of the data generating model under any of the simulation conditions. The $max(\hat{r}_{bis})$ measure only indicated this once under the conditions when $\sigma_\gamma^2 = 0.50$, 0.75, and 1.

### 5.5.1.2 Detection Rates Under the Posterior Predictive Method

Here we investigate the ability of each measure to indicate a lack of fit when the 2PN IRT model was fitted to data generated under the 2PN testlet model. The same steps were taken in conducting this portion of the simulation study as described in the above section 5.5.1.1. The exception being that now under each iteration we are fitting the 2PN IRT model to the observed testlet data $Y_{obs}$ and computing the posterior predictive samples of the measures from this. Table 5.5 gives the detection rates, or number of times out of 100 simulations each of the measures indicated a lack of fit with a $p$-value greater than 0.95. We see that under all four simulation conditions the detection rates are low.

As was seen in section 5.2.2, the larger the data generating value of $\sigma_\gamma^2$ the more discrepancy there is between posteriors from the fitted 2PN IRT and testlet model, particularly for
<table>
<thead>
<tr>
<th>( \sigma^2_\gamma )</th>
<th>( sd(\hat{r}_{bis}) )</th>
<th>( max(\hat{r}_{bis}) )</th>
<th>( mean(\hat{r}_{bis}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.50</td>
<td>1.0</td>
<td>4.0</td>
<td>0.0</td>
</tr>
<tr>
<td>0.75</td>
<td>9.0</td>
<td>13.0</td>
<td>0.0</td>
</tr>
<tr>
<td>1</td>
<td>33.0</td>
<td>27.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 5.5 Detection rates, or number of times out of 100 simulations that the 3 discrepancy measures \( sd(\hat{r}_{bis}) \), \( max(\hat{r}_{bis}) \), and \( mean(\hat{r}_{bis}) \), evaluated using the PP method, indicated a lack of fit with a \( p \)-value > 0.95. Response data \( Y \) was generated from a 2PN testlet model under 4 simulation conditions with testlet variance \( \sigma^2_\gamma = 0, 0.50, 0.75, \) and \( 1 \), and the fitted model was the 2PN IRT model.

the item discrimination parameters. So, we would expect higher detection rates corresponding to larger values of \( \sigma^2_\gamma \) and this is what we see. The highest rates, when \( \sigma^2_\gamma = 1 \), are 33% and 27% for \( sd(\hat{r}_{bis}) \) and \( max(\hat{r}_{bis}) \), respectively. When \( \sigma^2_\gamma = 0.75 \), both of these measures only detected a lack of fit about 10% of the time, and when \( \sigma^2_\gamma = 0.50 \) they detect it less than 5% of the time. The \( mean(\hat{r}_{bis}) \) measure did not indicate a lack of fit for any value of \( \sigma^2_\gamma \) used.

This last part is not entirely surprising however. As we saw in the example of figure 5.6 of section 5.4.1, the observed value of the \( mean(\hat{r}_{bis}) \) measure was close to the center of the posterior predictive distribution. This type of behavior is probably why the \( mean(\hat{r}_{bis}) \) measure did not detect any lack of fit throughout the 100 simulations. The observed values of the \( sd(\hat{r}_{bis}) \) and \( max(\hat{r}_{bis}) \) measures, which were not significant in the example, were still seen to be in the left tails of the posterior predictive samples. So, as we see these two measures show some power in detecting the lack of fit of the 2PN IRT model in this study, but are still very weak.

### 5.5.2 PPPS Simulation

Here we will discuss the results of the simulation study using the PPPS method with the measures \( sd(a) \), \( max(a) \), and \( mean(a) \), which were discussed in section 5.4.1. The first
part of the study was to investigate the type I error rates of these measures when the data generating model is fit to the observed testlet data $Y_{obs}$. The second part is to examine the detection rates of each measure when the 2PN IRT model is fitted to $Y_{obs}$. We again do this under four simulation conditions where the data generating value of the testlet variance $\sigma^2_\gamma$ is 0, 0.50, 0.75, and 1.

Each part of the simulation study was conducted in a similar fashion, the only difference was the model being fitted to the testlet data $Y_{obs}$. That is, either the 2PN testlet model was fitted for the first part, or 2PN IRT model for the second part. Posterior samples of length 1,000 of the discrimination parameters were drawn and posterior samples of the measures $sd(a)$, $max(a)$, and $mean(a)$ were computed from these. The 5 sample quantiles 0.05, 0.25, 0.50, 0.75, and 0.95 were then computed on each sample.

We then simulated 100 prior predictive response data sets $Y_{rep}^{(r)}$, $r = 1, \ldots, 100$, under the 2PN testlet model for the first part, and 2PN IRT model for the second part. We then fit the respective data generating model to each $Y_{rep}^{(r)}$ using an MCMC procedure. For $r = 1, \ldots, 100$, a posterior sample of the discrimination parameters of length 1,000 was drawn, which gave us a $40 \times 1000$ matrix where each column gives us a vector $\tilde{\alpha}^{(l)} = (\tilde{a}_1^{(l)}, \ldots, \tilde{a}_{40}^{(l)})^t$, $l = 1, \ldots, 1000$. We computed the standard deviation, maximum, and mean over each column which gave us a posterior sample of length 1,000 of each measure $sd(a)$, $max(a)$, and $mean(a)$. The 5 quantiles were then computed for each sample, giving a set of 100 posterior sample quantiles for each measure.

For each measure, the 100 Euclidean distances were then calculated between these 100 posterior sample quantiles and the computed mean quantile vector. The Euclidean distance $e_0$ for the posterior sample quantiles computed under $Y_{obs}$ is also found. Lastly, we compute a $p$-value by looking at the overall proportion of Euclidean distances greater than $e_0$ (see section 3.3.2). A $p$-value less than 0.05 is taken to indicate a significant lack of fit of the model to the data. For the first and second part, this total process was iterated 100 times to give a total of 100 $p$-values.
5.5.2.1 Type I Error Rates Under the PPPS Method

Here we show the number of times out of 100 simulations each of the three measures $sd(a)$, $max(a)$, and $mean(a)$ indicated a lack of fit of the 2PN testlet model to $Y_{obs}$. Table 5.6 gives the type I error rates for the measures where each row corresponds to the given data generating value of $\sigma_{\gamma}^2$. We see that nearly all the rates are at 5% or below. The exception being for the measure $mean(a)$ when $\sigma_{\gamma}^2$ was 0.50 or 0.75. When $\sigma_{\gamma}^2 = 0.50$, this measure indicated a lack of fit with the data generating model 6% of the time and when $\sigma_{\gamma}^2 = 0.75$ it indicated a lack of fit 7% of the time.

<table>
<thead>
<tr>
<th>Testlet Variance</th>
<th>Measures</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_{\gamma}^2$</td>
<td>$sd(a)$</td>
<td>$max(a)$</td>
<td>$mean(a)$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>4.0</td>
<td>4.0</td>
<td>5.0</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>3.0</td>
<td>3.0</td>
<td>6.0</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>4.0</td>
<td>3.0</td>
<td>7.0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.0</td>
<td>2.0</td>
<td>2.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6 The number of times out of 100 simulations under the PPPS method the 3 discrepancy measures $sd(a)$, $max(a)$, and $mean(a)$ indicated a lack of fit of the data generating model to the observed data $Y_{obs}$ generated under the 2PN testlet model. The observed testlet data $Y_{obs}$ was generated under 4 simulation conditions with testlet variance $\sigma_{\gamma}^2 = 0$, 0.50, 0.75, and 1. Each row above gives the type I error rates of the measures obtained using observed data $Y_{obs}$ generated under the corresponding value of $\sigma_{\gamma}^2$. A lack of fit was indicated with a $p$-value $< 0.05$.

As far as type I error is concerned, it appears that $sd(a)$ and $max(a)$ are the better of the three measures. Over the different data generating values of $\sigma_{\gamma}^2$, these two measures have rates that do not exceed the threshold level of 5%. However, even though some of the type I error rates of $mean(a)$ are greater than the threshold level of 5%, they are not that much greater. We should also keep in mind that the total number of simulations performed was 100, and so the potential type I error rates are discretized to whole numbers between 0 and 100. That is, the difference between having a type I error rate of 5% or 6% comes down to 1 simulation. If the number of simulations could be extended, as may be done in the future,
we may see these rates get closer to 5%.

5.5.2.2 Detection Rates Under the PPPS Method

In this part of the study we examine the abilities of the measures $sd(a)$, $max(a)$, and $mean(a)$ to detect a lack of fit of the 2PN IRT model to the observed testlet data $Y_{obs}$. Table 5.7 lists the number of times out of 100 simulations each measure indicated a lack of fit, where each row corresponds to the indicated value of $\sigma^2_\gamma$ used in simulating the observed testlet data sets.

<table>
<thead>
<tr>
<th>Detection Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testlet Variance $\sigma^2_\gamma$</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0.50</td>
</tr>
<tr>
<td>0.75</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.7 The number of times out of 100 simulations under the PPPS method the 3 discrepancy measures $sd(a)$, $max(a)$, and $mean(a)$ indicated a lack of fit of the 2PN IRT model to observed data $Y_{obs}$ generated under the 2PN testlet model. The observed testlet data $Y_{obs}$ was generated under 4 simulation conditions with testlet variance $\sigma^2_\gamma = 0, 0.50, 0.75, and 1$. Each row above gives the detection rates of the measures obtained using observed data $Y_{obs}$ generated under the corresponding value of $\sigma^2_\gamma$. A lack of fit was indicated with a $p$-value < 0.05.

Just as we saw with the detection rates of the measures under the PP method above in section 5.5.1.2, the detection rates for the measures increase as the testlet variance $\sigma^2_\gamma$ increases. When $\sigma^2_\gamma = 0$ we see that the detection rates for each measure is below 5%, which we would hope to see. When $\sigma^2_\gamma = 0.50$ the best performing measure is $sd(a)$ with a detection rate of 34% followed by $mean(a)$ with 29%. The $max(a)$ measure had the lowest rate at 21%.

When $\sigma^2_\gamma = 0.75$ we see that $mean(a)$ had the highest detection rate at 44% followed by $sd(a)$ with 41%. The measure $max(a)$ had the lowest rate here with 39%, although this is
not far behind the others. Lastly, when $\sigma_\gamma^2 = 1$ $sd(a)$ performed the best detecting a lack of fit a respectable 76% of the time. The other two measures’ rates were close with $mean(a)$ having 60% and $max(a)$ having 57%.

Looking at the performance of the three measures over the four simulation conditions, we see that the $sd(a)$ and $mean(a)$ measures perform the best in detecting a lack of fit of the 2PN IRT model. However, none of the measures were able to detect a lack of fit of the 2PN IRT model a great amount when $\sigma_\gamma^2$ was less than 1. If we wanted to indicate the best measure to use under the PPPS method we also consider the type I error rates that were found. Based on these, we would say $sd(a)$ is the better measure to use since all its type I error rates were under 5%.

5.6 Comparison of PP and PPPS Methods

We now take a look at the effectiveness of the two model checking procedures, given the measures we have used, in detecting the lack of fit of the 2PN IRT model to 2PN testlet data. As we saw, the posterior predictive method was not that powerful in detecting the lack of fit for any value of $\sigma_\gamma^2$ used. For the largest value of $\sigma_\gamma^2 = 1$, the $sd(\hat{r}_{bis})$ measure indicated a lack of fit only 33% of the time. Comparing results for each value of $\sigma_\gamma^2$ used, the PPPS method is much more powerful. Under the PPPS method the type I error rates of the measures were all under the threshold level of 5%, except for the $mean(a)$ measure when $\sigma_\gamma^2$ was 0.50 and 0.75, but they were not too bad.

So, given the measures studied, the PPPS method performs better than the posterior predictive method under these simulation conditions. However, as we can see in the results, our faith in the procedure to detect the lack of fit of the 2PN IRT model is still relatively weak for values of $\sigma_\gamma^2$ less than one. As mentioned, the current limitation of the PPPS method is the length of time it takes to complete. Especially when conducting simulation studies such as we have done here.

In our simulation study for the PPPS method, we simulated 100 separate prior predictive
response data sets under either the 2PN testlet model in the type I error rate study or the 2PN IRT model in the detection rate study. From these we simulated 100 posterior samples of each measure to compare to the posterior of the measure computed from $Y_{obs}$. As we have discussed, this is a low amount. Dey et al. (1998) suggested a good amount of separate prior predictive posteriors to be simulated was at least 500. Based on current computing restrictions we chose to use 100, but anticipate in future research more may be able to be used. In this, we may expect to see improved type I error and detection rates.
REFERENCES


Pearson, K. (1909). On a new method of determining correlation between a measured character a, and a character b, of which only the percentage of cases wherein b exceeds (or falls short of) a given intensity is recorded for each grade of a. *Biometrika* 7(1/2), 96–105.


