Posterior Predictive Assessment of Item Response Theory Models
Sandip Sinharay, Matthew S. Johnson and Hal S. Stern

Applied Psychological Measurement 2006 30: 298
DOI: 10.1177/0146621605285517

The online version of this article can be found at:
http://apm.sagepub.com/content/30/4/298

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for Applied Psychological Measurement can be found at:

- Email Alerts: http://apm.sagepub.com/cgi/alerts
- Subscriptions: http://apm.sagepub.com/subscriptions
- Reprints: http://www.sagepub.com/journalsReprints.nav
- Permissions: http://www.sagepub.com/journalsPermissions.nav
- Citations: http://apm.sagepub.com/content/30/4/298.refs.html

>> Version of Record - Jun 16, 2006

What is This?
Posterior Predictive Assessment of Item Response Theory Models

Sandip Sinharay, Educational Testing Service
Matthew S. Johnson, Baruch College
Hal S. Stern, University of California, Irvine

Model checking in item response theory (IRT) is an underdeveloped area. There is no universally accepted tool for checking IRT models. The posterior predictive model-checking method is a popular Bayesian model-checking tool because it has intuitive appeal, is simple to apply, has a strong theoretical basis, and can provide graphical or numerical evidence about model misfit. An important issue with the application of the posterior predictive model-checking method is the choice of a discrepancy measure (which plays a role like that of a test statistic in traditional hypothesis tests). This article examines the performance of a number of discrepancy measures for assessing different aspects of fit of the common IRT models and makes specific recommendations about what measures are most useful in assessing model fit. Graphical summaries of model-checking results are demonstrated to provide useful insights about model fit. 

*Index terms: Bayesian methods, discrepancy measures, model checking, odds ratio, p values*

Introduction

Model checking remains a major hurdle to overcome for effective implementation of item response theory (Hambleton & Han, 2004). The possible number of response patterns \(2^I\) for a test with \(I\) binary items) is large for even moderately long assessments, leading to sparse contingency tables—so the standard chi-square tests do not apply directly. Therefore, investigators assess the fit of item response theory (IRT) models by examining whether the model can explain various summaries (or collapsed versions) of the original data. Researchers van der Linden and Hambleton (1997) commented,

> The approach to the assessment of fit with two- and three-parameter models often involves the collection of a wide variety of evidence about model fit, including statistical tests, and then making an informed judgment about model fit and usefulness of a model with a particular set of data. (p. 16)

Even though a significant amount of research has concentrated on assessing different aspects of model fit in IRT, there is no universally accepted model fit measure (recent work by Maydeu-Olivares & Joe, 2005 notwithstanding). Standard IRT software packages are lacking in reliable model-fit indices, and worse, a number of IRT software packages produce model-fit indices that have well-known shortcomings (Hambleton & Han, 2004).
The situation is hardly better for IRT modeling under the Bayesian framework. There has been a recent surge in the use of Bayesian estimation in IRT. Patz and Junker (1999), Bradlow, Wainer, and Wang (1999), Beguin and Glas (2001), and Karabatsos and Sheu (2004) are only some of the recent examples applying Bayesian estimation and Markov chain Monte Carlo (MCMC) algorithms (e.g., Gelman, Carlin, Stern, & Rubin, 2003) to fit complicated psychometric models. However, not much attention has been given to assessing the fit of IRT models from a Bayesian perspective.

The posterior predictive model-checking (PPMC) method (Rubin, 1984) is a popular Bayesian model diagnostic tool because it has intuitive appeal, is simple to apply, has a strong theoretical basis, and can provide graphical or numerical evidence about model misfit. The method compares the observed data with replicated data (data that are generated or predicted by the model) using a number of diagnostic measures that are sensitive to model misfit. Any systematic differences between aspects of the observed data set and those of the replicated data sets indicate a failure of the model to explain those aspects of the data. Graphical display is the most natural and easily comprehensible way to perform posterior predictive checks (PPC); in situations where graphical displays do not suffice or are cumbersome (e.g., for too many checks simultaneously), one can also use a tail-area probability, also known as a posterior predictive p-value (PPP-value). Examples of application of the PPMC method in educational measurement include Hoijtink (2001), Beguin and Glas (2001), and Karabatsos and Sheu (2004). However, most of these applications are specialized, and few of these focus in general on unidimensional IRT models. Before applying any technique to complicated models in a field, it is important to study its performance with simpler models (especially in IRT, where even the basic models are not so simple as far as model checking is concerned). However, attempts to thoroughly study model checking have been rare in the context of Bayesian IRT.

Motivated by the aforementioned, this article uses the PPMC method in a number of simulations and in a real data example to assess several aspects of fit of unidimensional IRT models for dichotomous items. Use of simple and natural discrepancy measures reveals useful information about model misfit when model assumptions are violated. This article emphasizes graphical displays of PPCs. For some aspects of the model, PPCs have to be carried out for a large number of measures (e.g., for all pairs of items). This article develops graphical approaches for summarizing the output of such high-dimensional model checks in a precise and easily comprehensible manner. The PPMC method has previously been applied to examine person fit, item fit (Sinharay, in press) and differential item functioning (Hoijtink, 2001) for the simple IRT models—so these aspects are not covered in this article.

It is important to emphasize that model checking is related to but different from model selection. The goal in model checking is to determine if the model is adequate or not. A model that is found to fit the data (or, more precisely, is not found to not fit the data) is not necessarily the best model or the correct model. In fact, there may be several models that might fit the data in the sense of being found acceptable by model checking. Although it is important to be aware of the possibility that other models may fit the data (see, e.g., Bickel, Buyske, Chang, & Ying, 2001), it remains important to assess whether one’s chosen model is appropriate for analyzing the data at hand.

The organization for the remainder of this article is as follows. The first section introduces the PPMC method and discusses interpretation of model checks. The second section discusses the discrepancy measures examined in this study. The third section describes an outline of the simulation studies. The fourth section describes the results of the simulations, and the next section uses those results to examine the fit of the Rasch model to data from the U.S. Department of Agriculture’s (USDA’s) Measurement of Food Security program. The final section summarizes the work and provides the reader with recommendations about how to use posterior predictive checks in IRT.
Posterior Predictive Model-Checking (PPMC) Techniques

Description of the Method

Let \( p(y|\omega) \) denote the likelihood for a statistical model applied to data \( y \) (examinee responses in this context), and let \( p(\omega) \) be the prior distribution on the parameters, where \( \omega \) denotes all the parameters in the model. Then the posterior distribution of \( \omega \) is \( p(\omega|y) \propto p(y|\omega)p(\omega) \). The PPMC method (Rubin, 1984) suggests checking a model by examining whether the observed data \( y \) appear extreme with respect to the posterior predictive distribution of replicated data \( y^{rep} \),

\[
p(y^{rep}|y) = \int p(y^{rep}|\omega)p(\omega|y)d\omega. \tag{1}
\]

In practice, test quantities or discrepancy measures \( D(y, \omega) \) are defined (Gelman, Meng, & Stern, 1996), and the posterior distribution of \( D(y, \omega) \) is compared to the posterior predictive distribution (PPD) of \( D(y^{rep}, \omega) \), with substantial differences between them indicating model misfit. A researcher may use \( D(y, \omega) = D(y) \), a discrepancy measure depending on the data only (which can also be called a test statistic); in that case, the PPC consists of comparing \( D(y) \) to the PPD of \( D(y^{rep}) \). The PPMC method operates in the spirit of a diagnostic plot rather than as a hypothesis test (Stern, 2000), and thus the comparison of observed and replicated discrepancy measures is mostly performed using graphical plots. In cases where graphical checks are not conclusive or when one is interested in several checks simultaneously, it can be useful to examine a quantitative measure of lack of fit, a tail-area probability also known as the PPP-value:

\[
P(D(y^{rep}, \omega) \geq D(y, \omega)|y) = \int_{D(y^{rep}, \omega)\geq D(y, \omega)} p(y^{rep}|\omega)p(\omega|y)dy^{rep}d\omega, \tag{2}
\]

which can be a useful supplement to a graphical plot.

Because of the difficulty in dealing with equation (1) or (2) analytically for all but simple problems, Rubin (1984) suggested simulating replicated (or posterior predictive) data sets from the PPD in applications of the PPMC method. One draws \( N \) simulations \( \omega^1, \omega^2, \ldots, \omega^N \) from the posterior distribution \( p(\omega|y) \) of \( \omega \) (most likely using an MCMC algorithm) and then draws \( y^{rep,n} \) from the distribution \( p(y^{rep}|\omega^n) \) for \( n = 1, 2, \ldots N \). The process results in \( N \) draws from the joint posterior distribution \( p(y^{rep}, \omega|y) \). One then computes the predictive discrepancies \( D(y^{rep,n}, \omega^n) \) and realized discrepancies \( D(y, \omega^n) \), \( n = 1, 2, \ldots N \). It is possible then to create a graphical plot of \( D(y^{rep,n}, \omega^n) \) versus \( D(y, \omega^n) \), \( n = 1, 2, \ldots N \); points lying consistently above or below the 45-degree line indicate model misfit. The proportion of the \( N \) replications for which \( D(y^{rep,n}, \omega^n) \) exceeds \( D(y, \omega^n) \) provides an estimate of the PPP-value. Extreme PPP-values (close to 0, 1, or both, depending on the nature of the discrepancy measure) indicate model misfit. Figure 1 graphically describes the PPMC method.

Interpreting Posterior Predictive Model Diagnostics

Posterior predictive (PP) model checks are offered in the spirit of model diagnostics such as residual plots in regression. Plots suggest when the observed data do not look like replicated data under the model. Investigators can decide if the difference they see seems important enough to warrant fitting a new model (which is often suggested by the plots; Gelman et al., 2003, p. 176) or taking other remedial action.

Graphical displays are ambiguous on occasion; some other times, individuals prefer to make a judgment based on a quantitative measure. The PPP-values can be useful in this regard—they provide a quantitative measure of the degree to which observed data such as the ones at hand would be...
expected under the model. Extremely small probabilities (less than .01 or greater than .99) are very suggestive that the model does not capture the corresponding feature of the data. Less extreme values (say less than 0.05) are also suggestive of model failure, although as described in the next paragraph, such a threshold should not be viewed as a significance level in the usual sense.

Gelman et al. (1996) commented that they “use the term assessment instead of testing to highlight the fundamental difference between assessing the discrepancies between a model and data and testing the correctness of a model” (p. 734). Thus, the PPP-values do not refer to an explicit hypothesis test. The PPP-values share some features of traditional hypothesis-testing p values; in particular, both are defined as tail-area probabilities, but there are important differences as well. The PP approach does not carry out a hypothesis test, and the PPP-values are not necessarily uniformly distributed when the fitted model is in fact correct, and there is some evidence that PPP-values under the correct model tend to be closer to 0.5 more often than would be expected under a uniform distribution (Bayarri & Berger, 2000; Sinharay & Stern, 2003). One purpose of the simulation studies is to investigate the properties of the PP model diagnostics (including PPP-values) when they are applied to IRT models.

**Application to Item Response Theory Models**

The approach as it applies in the context of IRT models is described next. Let \( y_{ij} \) denote the binary score of the \( i \)th individual for the \( j \)th item in an educational assessment. Suppose that the IRT model of interest is the three-parameter logistic (3PL) model,

\[
p_{ij} = P(y_{ij} = 1) = c_j + (1 - c_j)\left\{1 + \exp(-a_j(\theta_i - b_j))\right\}^{-1},
\]

with symbols having their usual meanings. In this context, if the examinee proficiencies \( \theta_i \) are treated as nuisance parameters, then \( \omega \) is the collection of all item parameters. The posterior distribution \( p(\omega|y) \) is given by

\[
p(\omega|y) \propto \text{likelihood} \times \text{prior} = \left[ \prod_i \int_{\theta_i} \left\{ \prod_j p_{yij}(1 - p_{ij})^{1-y_{ij}} \right\} p(\theta_i) d\theta_i \right] p(\omega),
\]

where \( p(\theta_i) \) is the population distribution on the \( \theta_i \). To assess the fit of the 3PL model to a data set, one has to repeat the following steps a large number of times:
1. Generate a draw of item parameters from the posterior distribution given by equation (3) (using, as in Patz & Junker, 1999, an MCMC algorithm to fit the 3PL model). Draw proficiency parameters \( \theta_i \)s from \( p(\theta) \).

2. Draw a data set from the 3PL model, using the item parameters and proficiency parameters drawn in the above step.

3. Compute the values of the predictive and realized discrepancy measures from the above draws of parameters and data set.

One then creates plots and/or computes PPP-values for the discrepancy measures.

### Discrepancy Measures

The choice of discrepancy measures is crucial in the application of the PPMC method. Although the flexibility of the method allows one to use any function of data and parameters as a discrepancy measure, all measures may not be useful. Ideally, discrepancy measures should be chosen to reflect aspects of the model that are relevant to the scientific purposes to which the inference will be applied and to measure features of the data not directly addressed by the probability model (Gelman et al., 2003, p. 172). It is easy to demonstrate that the observed proportions correct for the items are not helpful diagnostics because these quantities will be fit well even by an inappropriate IRT model as long as the model includes difficulty parameters that relate closely to the proportions correct. It was also found that the chi-square-type measure \( \sum_j \frac{(\hat{y}_{ij} - E(y_{ij}))^2}{\text{Var}(y_{ij})} \), the sum of squares of the outfit measures (e.g., van der Linden & Hambleton, 1997, p. 113), although intuitive and useful for other types of statistical models, is not useful for IRT model checking as it fails to detect problems with inadequate models. This last result suggests caution; before applying the PPMC method with a particular discrepancy measure, one should examine how useful that measure is for detecting failures of IRT models.

For unidimensional IRT models, a number of discrepancy measures may be of interest, depending on the context of the problem. These models assume unidimensionality, local independence (LI), a specific shape of the response function, and, often, normality of the ability distribution, and each of these assumptions should be checked using suitable discrepancy measures. There does not seem to exist a single, omnibus discrepancy measure that can detect violations of all of these assumptions together. In the sections that follow, this article shows how different discrepancy measures can be used to detect various model violations.

### Discrepancy Measures Examined

#### Observed Score Distribution

Let \( NC_j \) denote the number of examinees getting exactly \( j \) items correct, \( j = 0, 1, 2, \ldots J \). Furthermore, let \( NC = (NC_0, NC_1, \ldots NC_J) \). Ferrando and Lorenzo-seva (2001) and Hambleton and Han (2004) suggested comparing the observed and predicted score distributions to measure overall model fit. To summarize the model fit to the observed score distribution in a single number, Beguin and Glas (2001) suggested the discrepancy measure \( \chi^2_{NC} = \sum_j \frac{(NC_j - E(NC_j))^2}{E(NC_j)} \), where \( E(NC_j) \) is the expectation of \( NC_j \). Although the statistic \( \chi^2_{NC} \) does not follow a chi-square distribution, the PPMC method automatically provides the relevant reference distribution.
Biserial Correlation Coefficient

The correlation of examinee scores with the binary outcomes on a particular item, the biserial correlation coefficient, may be appropriate for detecting the inadequacy of IRT models that do not include a slope parameter, as shown in a small example in Albert and Ghosh (2000), where the Rasch model was found inadequate.

Measures of Association Among the Item Pairs

Unidimensional IRT models imply specific types of associations between the responses of an individual to pairs of items. So, discrepancy measures that examine the association between item pairs can potentially detect lack of fit of such models. This article examines two such statistics.

Odds ratio (OR). Consider an item pair. Let \( n_{kk'} \) denote the number of individuals scoring \( k \) on the first item and \( k' \) on the second item, \( k, k' = 0, 1 \). This study uses \( OR = \frac{n_{11}}{n_{00}} \), the sample odds ratio (e.g., Agresti, 2002, p. 45), referred to as the odds ratio (OR) hereafter, as a test statistic with the PPMC method. Chen and Thissen (1997) used the OR to detect violations of the LI assumption in unidimensional IRT models; they argued that if LI is satisfied, a unidimensional model can fit the observed OR. However, LI may not hold for a particular application for reasons such as speediness or the test not being unidimensional in a psychological sense (e.g., Chen & Thissen, 1997); in that case, the observed OR will be larger (smaller) than what is expected under a unidimensional IRT model for within-cluster items (between-cluster items). Chen and Thissen found the standardized log-OR not to have a \( N(0, 1) \) null distribution and hence did not find it a useful diagnostic. The PPMC method overcomes this problem because it does not require an explicit distributional assumption for the discrepancy measure.

Mantel-Haenszel (MH) statistic. Let us define the sample odds ratio for an item pair conditional on the rest score (i.e., the raw score on the test obtained by excluding the two items) \( r \) as \( OR_r = \frac{n_{11}r_{00}}{n_{10}r_{01}} \), where \( n_{kk'r} \) is the number of individuals with rest score \( r \) obtaining a score \( k \) on the first item and \( k' \) on the second, \( k, k' = 0, 1 \). Then the \( OR_r \)s can be combined into an MH statistic (e.g., Agresti, 2002, p. 234), as \( MH = \sum_{r} \frac{n_{11}r_{00}/n_r}{n_{10}r_{01}/n_r} \), where \( n_r \) is the number of examinees obtaining rest score \( r \). If LI holds, using arguments as in Stout et al. (1996), the conditional covariance between the scores on the two items is close to zero, and the MH statistic should be near 1; if LI is violated, the conditional covariance is positive (which means the MH statistic is more than 1) for within-cluster items and negative (MH statistic less than 1) for between-cluster items. Hence, as with the OR, in situations where LI is violated, the observed MH statistics are likely to be higher (lower) than expected for within-cluster (between-cluster) item pairs.

Outline of the Simulation Studies

To explore the performance of the PPMC method in detecting lack of fit for IRT models, we apply the method to a variety of combinations of the data-generating model and data analysis model. The combinations in Table 1 with an asterisk (“*”) in the corresponding cells are the cases examined. More details about the two-dimensional model and the “3PL, nonnormal \( p(\theta) \)” model can be found in the sections where these are discussed. The simulation approach in this study is as follows.

Consider a data-generating model \( M_g \) and a data analysis model \( M_a \). One hundred data sets are generated from \( M_g \), each data set containing responses of 2,500 examinees to 30 items. The generating parameter values (for items and examinees) are the same for all the 100 data sets generated.
The proficiency parameters are generated from a $N(0,1)$ distribution. The item parameters used in $M_g$ are similar to those encountered in real test data. For example, when $M_g$ is the 3PL model, this work uses real item parameter estimates, shown in Table 2, from the National Assessment in Educational Progress (NAEP). For generating data from the Rasch or two-parameter logistic (2PL) model, relevant parameters from the same table are used. For each of the 100 data sets generated, the model $M_a$ is fit using an MCMC algorithm (five chains of length 6,000 are run, which ensures convergence of the MCMC algorithm; the first 2,000 draws in each chain are discarded; and the remaining output is thinned by retaining every 20th draw to get a combined posterior sample of 1,000). The prior distributions used for fitting the models are

$$
\log(a_j) \sim \mathcal{N}(0, 10), \quad b_j \sim \mathcal{N}(0, 10), \quad \text{logit}(c_j) \sim \mathcal{N}(\logit(0.2) = -1.4, 1).
$$

The population distribution of proficiencies is taken to be $\theta_i \sim N(0, 1)$. The results of the PPCs in the simulation study are robust to reasonable changes to the prior distributions (as long as prior variances are not close to zero).

### Table 1
The Cases Examined in the Simulation Studies

<table>
<thead>
<tr>
<th>Analysis Model</th>
<th>1PL</th>
<th>2PL</th>
<th>3PL</th>
<th>Two-Dimensional</th>
<th>3PL, Nonnormal $p(\theta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1PL</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2PL</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3PL</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* 1PL = one-parameter logistic; 2PL = two-parameter logistic; 3PL = three-parameter logistic.

### Table 2
The Generating Parameter Values for the Simulation Studies

<table>
<thead>
<tr>
<th>Item ID</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>$a_j$</td>
<td>0.70</td>
<td>1.89</td>
<td>1.32</td>
<td>1.36</td>
<td>1.17</td>
<td>0.56</td>
<td>1.10</td>
<td>1.68</td>
<td>1.01</td>
<td>0.88</td>
</tr>
<tr>
<td>$b_j$</td>
<td>1.81</td>
<td>-0.52</td>
<td>0.26</td>
<td>-1.48</td>
<td>-0.52</td>
<td>0.44</td>
<td>-2.15</td>
<td>0.96</td>
<td>-0.87</td>
<td>1.70</td>
</tr>
<tr>
<td>$c_j$</td>
<td>0.08</td>
<td>0.66</td>
<td>0.14</td>
<td>0.17</td>
<td>0.17</td>
<td>0.14</td>
<td>0.30</td>
<td>0.25</td>
<td>0.10</td>
<td>0.22</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item ID</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_j$</td>
<td>1.19</td>
<td>0.60</td>
<td>1.49</td>
<td>2.01</td>
<td>2.40</td>
<td>2.00</td>
<td>1.48</td>
<td>1.10</td>
<td>1.52</td>
<td>1.45</td>
</tr>
<tr>
<td>$b_j$</td>
<td>-2.41</td>
<td>-0.56</td>
<td>-0.27</td>
<td>0.04</td>
<td>-0.79</td>
<td>-0.38</td>
<td>0.14</td>
<td>-1.53</td>
<td>0.11</td>
<td>-0.21</td>
</tr>
<tr>
<td>$c_j$</td>
<td>0.23</td>
<td>0.21</td>
<td>0.12</td>
<td>0.20</td>
<td>0.18</td>
<td>0.17</td>
<td>0.13</td>
<td>0.14</td>
<td>0.20</td>
<td>0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item ID</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_j$</td>
<td>0.80</td>
<td>0.67</td>
<td>0.83</td>
<td>1.17</td>
<td>1.43</td>
<td>2.4</td>
<td>1.53</td>
<td>1.2</td>
<td>1.7</td>
<td>2.05</td>
</tr>
<tr>
<td>$b_j$</td>
<td>-0.52</td>
<td>-0.43</td>
<td>-1.25</td>
<td>-0.52</td>
<td>-0.27</td>
<td>-0.44</td>
<td>1.75</td>
<td>-0.80</td>
<td>0.40</td>
<td>-0.93</td>
</tr>
<tr>
<td>$c_j$</td>
<td>0.11</td>
<td>0.12</td>
<td>0.18</td>
<td>0.13</td>
<td>0.15</td>
<td>0.40</td>
<td>0.25</td>
<td>0.24</td>
<td>0.16</td>
<td>0.26</td>
</tr>
</tbody>
</table>

The proficiency parameters are generated from a $N(0,1)$ distribution. The item parameters used in $M_g$ are similar to those encountered in real test data. For example, when $M_g$ is the 3PL model, this work uses real item parameter estimates, shown in Table 2, from the National Assessment in Educational Progress (NAEP). For generating data from the Rasch or two-parameter logistic (2PL) model, relevant parameters from the same table are used. For each of the 100 data sets generated, the model $M_a$ is fit using an MCMC algorithm (five chains of length 6,000 are run, which ensures convergence of the MCMC algorithm; the first 2,000 draws in each chain are discarded; and the remaining output is thinned by retaining every 20th draw to get a combined posterior sample of 1,000). The prior distributions used for fitting the models are

$$
\log(a_j) \sim \mathcal{N}(0, 10), \quad b_j \sim \mathcal{N}(0, 10), \quad \text{logit}(c_j) \sim \mathcal{N}(\logit(0.2) = -1.4, 1).
$$

The population distribution of proficiencies is taken to be $\theta_i \sim N(0, 1)$. The results of the PPCs in the simulation study are robust to reasonable changes to the prior distributions (as long as prior variances are not close to zero).
The posterior samples are used to perform the PPCs. As a numerical summary, the PPP-value for each discrepancy measure is computed on each data set, resulting in 100 PPP-values for each discrepancy measure. Note that the PPP-values are computed here to allow efficient summarization of the results of the many model-checking simulations; in a real application, a graphical plot is recommended first and PPP-values, if required, later.

Results From the Simulation Studies

Analyses when the fitted model is the same as the generating model are useful for learning about the frequency with which misfit is wrongly detected by an approach. In this study’s case, such simulations suggest that PPCs are, if anything, conservative in that they tend not to show misfit of a correct model too often. For example, only 3% of the ORs had PPP-values less than .05 when the 3PL model was used to both generate and analyze the data.

The remainder of the discussion is concentrated on cases where the generating model and analysis model are different. Findings that indicate how different discrepancy measures help detect different types of lack of fit are emphasized. It is true that the correct model is known, that Bayarri and Berger (2000) recommended that it is more appropriate to use model selection tools (Bayes factor, Bayesian information criterion [BIC], etc.) than model checks if an investigator has a clearly formulated alternative to the hypothesized model \( H_0 \), and that Gelman et al. (1996) recommended the PPMC method for assessing the fit of a single model to the available data in the absence of an alternative model. However, this study’s approach best reflects what happens in operational settings, where one often has a single preferred model at his or her disposal. Most of the operational tests using IRT employ a predetermined model, and there is no alternative model (e.g., the NAEP employs the 3PL model for dichotomous item calibration); if the model does not fit the data, analysts often employ remedial actions such as discarding items or collapsing item responses until the model fits the data. The advantage of generating from known models in these simulations, other than convenience, is that it allows one to know the type of misfit to expect so that it is possible to cross-check how the PPMC method has fared.

In a typical real application, an investigator will analyze one data set; so for most of the simulation scenarios, results for a single simulated data set appear first. However, it is also a goal to learn how the PPMC method performs in repeated analyses (Rubin, 1984); so, results for 100 data sets under each simulation scenario are also given. To compare results to those when the model fits, the plots under the true model (i.e., when the analysis model is the same as the data-generating model) are also shown in most of the figures.

Data-Generating Model: 2PL/3PL; Data Analysis Model: 1PL

The one-parameter logistic (1PL)/Rasch model is the simplest of the IRT models and makes restrictive assumptions (equality of item discriminations and no guessing); discrepancy measures that can exploit the limitations, if any, of these assumptions are most likely to identify misfit.

Because the biserial correlation coefficient is closely related to the item slope parameter, data generated from the 2PL/3PL model are likely to have more extreme biserial correlations (both high and low) than can be explained by the Rasch model.

The top left panel of Figure 2 shows, for one data set generated from the 3PL model, the observed biserial correlations, corresponding 90% PP intervals, and PP medians. Only a few of the PP intervals contain the observed biserial coefficients, indicating the success of the PPMC method in detecting the lack of fit of the Rasch model to the biserial correlations. The replicated medians
are all very similar, around 0.57. This is not surprising as the Rasch model assumes that all items have the same slope parameter. The bottom left panel shows that these patterns are not seen when the correct model is used to analyze the simulated data. It is straightforward to construct a single summary measure involving the biserial correlations that demonstrates the lack of fit of the Rasch model. The top right panel of Figure 2 shows, as in Albert and Ghosh (2000), the observed and 1,000 replicated values of the standard deviation (SD) of the biserial correlations. Here, the huge difference of the observed value (0.13) from the replicated values (replicated median is only about 0.025), as against no difference for the bottom right panel of the figure, points to the inadequacy of the Rasch model for this data set. The same phenomenon occurs in the other 99 data sets as well (results not shown).

The different slope parameters in the 2PL/3PL model induce a degree of dependence among the items (Chen & Thissen, 1997, p. 269); the Rasch model, with the same slope for all items, fails

Note. The top row shows plots for biserial correlations for a single data set generated from the three-parameter logistic (3PL) model and analyzed by the one-parameter logistic (1PL) model. The left-hand plot shows the observed biserial for each item as a solid ellipse, corresponding 90% posterior predictive (PP) intervals as a dotted line, and the PP median as a hollow ellipse; the right-hand plot shows a histogram as a summary of the PPD of the standard deviation (SD) of the biserial correlations. A dotted vertical line shows the observed SD. For comparison purposes, the bottom row shows similar plots when the same data are analyzed by the correct (3PL) model.
to explain the dependence. To illustrate, a case is considered in which the generating model is the 2PL model. Figure 3 plots the observed ORs involving Items 1 and 2, corresponding 90% PP intervals, and PP medians from the analysis of a single data set generated from the 2PL model. For example, for the item pair (1, 2), the observed OR is 1.8, whereas a corresponding 90% PP interval is (1.75, 2.46).

The figure reveals a number of interesting facts. First, the replicated medians of the ORs are near 2 for each item pair; this phenomenon has not been reported in the literature yet. Intuitively, the Rasch model assumes the same discrimination parameter for all the items and hence has all item response functions parallel to each other, causing the predicted ORs to be close to one another. Haberman, Holland, and Sinharay (in press), motivated by these findings, recently suggested some theoretical results regarding this phenomenon. Next, by comparing the observed ORs to the replicated ORs, it is obvious that the PPMC method succeeds in detecting the inadequacy of the Rasch model to capture the associations among the items in data sets that exhibit more complex structure.
The observed ORs for pairs involving Item 1 (an item with low discrimination) appear quite small compared to the reference distribution obtained under the Rasch model. The opposite is true for the ORs for pairs involving Item 2 (one with high discrimination). The plots in the bottom row, in sharp contrast, do not show any such patterns.

Figure 4 summarizes, from all 100 simulated data sets, PPP-values for the OR for each pair of items when data are generated from the 2PL model but analyzed with the Rasch model. The items are sorted in increasing order of their true discrimination parameters.

The plot is a variation of one used in Friendly (2002); it simultaneously displays two aspects of the simulation results. The circles below the main diagonal (the ones to the right of an imaginary 45-degree line through the middle of the plot) show the median of the 100 posterior predictive p values (PPP-values) for each item pair; the proportion of the circle that is shaded is equal to the median. In the same manner, the circles above the main diagonal show the proportion of simulated data sets with extreme PPP-values.

The observed ORs for pairs involving Item 1 (an item with low discrimination) appear quite small compared to the reference distribution obtained under the Rasch model. The opposite is true for the ORs for pairs involving Item 2 (one with high discrimination). The plots in the bottom row, in sharp contrast, do not show any such patterns.

Figure 4 summarizes, from all 100 simulated data sets, PPP-values for the OR for each pair of items when data are generated from the 2PL model but analyzed with the Rasch model. The items are sorted in increasing order of their true discrimination parameters.

The plot is a variation of one used in Friendly (2002); it simultaneously displays two aspects of the simulation results. The circles below the main diagonal (the ones to the right of an imaginary 45-degree line through the middle of the plot) show the median of the 100 PPP-values for each item pair; the proportion of a circle that is shaded is equal to the corresponding median PPP-value. Thus, item pair (1, 2) has a median PPP-value near 1, whereas item pair (29, 30) has a median PPP-value near zero. The circles above the main diagonal (those to the left of the imaginary 45-degree line) provide some information about how often the PPMC technique detects lack of fit for each item pair. The shaded proportion of a circle for an item pair is the same as the proportion

Note. The circles below the main diagonal show the median of 100 posterior predictive p values (PPP-values) for each item pair; the proportion of the circle that is shaded is equal to the median. In the same manner, the circles above the main diagonal show the proportion of simulated data sets with extreme PPP-values.
of times the PPP-values are extreme (below 0.05 or above 0.95) for that item pair. The two halves of Figure 4 tell essentially the same story—whenever there is an extreme median for an item pair (either extremely high or low), the proportion of extreme $p$ values is high for the same item pair. For comparison purposes, Figure 5 shows what happens when the correct model (i.e., the 2PL model) is fitted to the data. In Figure 5, the median PPP-values (below the diagonal) are all near 0.5, and the proportions of extreme PPP-values (above the diagonal) are small.

**Data-Generating Model: 3PL; Data Analysis Model: 2PL**

It proved relatively easy above to identify misfit of the Rasch model for data in which it was not appropriate. This next case is more difficult. The fit of the 2PL model has often been found adequate when data are generated from the 3PL model (see, e.g., Sinharay, in press).

Figure 6 compares, for a data set generated from the 3PL model, the observed number of examinees attaining a particular score to its PPD under the 2PL (top row) and 3PL (bottom row) models. The top left panel of Figure 6 (similar to a plot used by Béguin & Glas, 2001) shows that the observed score frequencies lie outside the 90% interval in a number of places. There are fewer people at the low end of the observed score distribution than expected under the 2PL model.

---

**Note.** The circles below the main diagonal show the median of 100 posterior predictive $p$ values (PPP-values) for each item pair; the proportion of the circle that is shaded is equal to the median. In the same manner, the circles above the main diagonal show the proportion of simulated data sets with extreme PPP-values.
because the 2PL model includes no guessing parameter. More (less) people fall in the 10 to 12 (22-25) range than expected under the 2PL model. The summary discrepancy measure $\chi^2_{NC}$ is very informative in this situation. The top right panel of Figure 6 (in contrast to the bottom right panel) shows that the realized values of $\chi^2_{NC}$ are mostly larger than the predictive values for this data set.

Figure 7 shows that the patterns observed above in the one data set persist across repetitions. Under the 2PL model, PPP-values are consistently high for low raw scores and those in the range of 22 to 25 correct, indicating that these raw scores occur less often than expected. The PPP-values are consistently low in the middle range and for the top scores, indicating that these raw scores occur more often than expected. The PPP-value for the aggregate measure $\chi^2_{NC}$ is less than 0.04 for each of the 100 generated data sets. Therefore, the PPMC method demonstrates that the 2PL model cannot adequately explain the observed score distribution of data from the 3PL model.

Biserial correlation is another aspect of data generated from the 3PL model that the 2PL model fails to explain adequately. For one data set generated from the 3PL model, for a number of items (mostly those with high biserials), the observed biserial correlations are lower than expected under

---

**Figure 6**

Fit of the 2PL Model and the 3PL Model to the Observed Score Distribution

---

*Note.* The top row shows plots of diagnostics based on the observed score distribution for a data set generated from the three-parameter logistic (3PL) model and analyzed by the two-parameter logistic (2PL) model. The left panel shows the observed score distribution (solid ellipses, connected by a solid line) and the corresponding posterior predictive distribution (PPD; hollow ellipses joined by dotted lines at posterior predictive [PP] medians and dotted lines indicating lower and upper bounds of PP 90% intervals); the right panel shows, along with a diagonal line, the realized and predictive discrepancies for $\chi^2_{NC}$. For comparison purposes, the bottom row shows similar plots when the same data are analyzed by the correct (3PL) model.
the 2PL model (although the discrepancy is not as extreme as observed for the Rasch model). As an outcome, the observed mean and $SD$ of the biserial correlations are considerably lower than expected under the 2PL model (see Figures 18 and 19 in Sinharay & Johnson, 2003). The same pattern persists across repetitions. Figure 8 shows histograms of the PPP-values for the mean and $SD$ of the biserial correlations across the 100 simulated data sets for the 2PL and 3PL models. The PPP-values for both mean and $SD$ for the 2PL model are concentrated on high values.

The finding that observed scores and biserial correlations can identify weaknesses in the fit of the 2PL model to 3PL data is important in that there has previously been little success distinguishing the 2PL and 3PL models.

**Data-Generating Model: Multidimensional; Data Analysis Model: 3PL**

To study the ability of the PPMC method to detect lack of fit when the local independence assumption of the unidimensional IRT models is violated, the case is considered in which the 3PL model is fit to data generated from the linear logistic 2-dimensional model (Reckase, 1997, and the references therein) given by

$$
P(y_{ij} = 1) = c_j + (1 - c_j)(1 + e^{-(a_1 j \theta_{1i} + a_2 j \theta_{2i} - b_j)})^{-1},
$$

$$(\theta_{1i}, \theta_{2i}) \sim N_2(0, 0, 1, 1, \rho), |\rho| \leq 1.$$  

**Note.** The left panel shows boxplots showing the 100 posterior predictive $p$ values (PPP-values) for each observed raw score across the 100 simulated data sets generated from the three-parameter logistic (3PL) model and analyzed by the two-parameter logistic (2PL) model. The width of the boxplot for any observed score is proportional to the mean number of people for that score. There are horizontal dotted lines at 0.05 and 0.95. For comparison purposes, the right panel shows similar plots when the same data sets are analyzed under the correct (3PL) model.
For generating data from the above model, the same difficulty and guessing parameters are used as in Table 2. The closer the value of \( \rho \) to 1, the more unidimensional are the data. The slope parameters for the two dimensions, \( a_{1j} \) and \( a_{2j} \), are functions of the slopes \( a_j \) used in Table 2, \( a_{1j} = a_j I_{j \leq 15} \), \( a_{2j} = a_j I_{j > 15} \). Under this data-generating model, the first and second halves in the test measure different correlated skills.

The following discussion describes the results for the case with \( \rho = 0.7 \). Association among the items is an aspect the unidimensional 3PL model should fail to explain here. The MH statistic detects model misfit more often than the OR and is featured here. Figure 9, using the same symbols as in Figure 4, shows the PPP-values for a single data set above the main diagonal and the median PPP-values for 100 data sets below the main diagonal.

There an abundance of extreme PPP-values. It is important to note here (and in other examples) that some extreme PPP-values are to be expected by chance even under a correct model given the large number of item pairs considered (435 in this case). One should interpret the plot as showing misfit only if the number of extreme \( p \) values is large and/or the extreme \( p \) values indicate an unusual pattern (e.g., all the extreme \( p \) values involving the same few items). Figure 9 clearly indicates that the 3PL model does not fit because there are many extreme PPP-values, and the extreme \( p \) values indicate that the items form two clusters. Observed associations (as measured by the MH statistic) for within-cluster\(^3\) (between-cluster) item pairs are mostly higher (lower) than expected under the 3PL model. This phenomenon is mentioned by, for instance, Stout et al. (1996), for multidimensional assessments, but Figure 9 provides an attractive graphical demonstration of the fact.

\[ \text{Note. The vertical dashed lines show the corresponding medians.} \]
The patterns of Figure 9 are not present when the appropriate two-dimensional model is fit to the data and a similar plot (not shown) created. It is also possible to detect misfit of the 3PL model using the OR statistic (see Sinharay & Johnson, 2003, for results) and more often using the MH statistic (results not shown) when (a) data are generated from the testlet model (Bradlow et al., 1999), a special case of a multidimensional model (Wang & Wilson, 2005, p. 129), and (b) when data are generated from a speeded test. For both these situations, the local independence assumption of a 3PL model is known to be violated.

Data-Generating Model: 3PL With a Nonnormal Ability Distribution; Data Analysis Model: 3PL With $\mathcal{N}(0, 1)$ Ability Distribution

Here, the generating item characteristic curves (ICCs) are those of a 3PL model, but the generating ability distribution, $p(\theta)$, is a mixture of two normal distributions, $\mathcal{N}(-0.253, 0.609^2)$ and $\mathcal{N}(2.192, 1.045^2)$, with mixing proportions 0.897 and 0.103, respectively, as in Woods and Thissen (2004). As a result, $p(\theta)$ has mean 0 and SD 1, as in the $\mathcal{N}(0, 1)$ distribution, but has skewness 1.9 and kurtosis 5. A simple 3PL model using the $\mathcal{N}(0, 1)$ population distribution assumption is fit to the data. The observed score distribution, which is closely related to the underlying ability distribution, is an aspect of the data that the fitted model fails to describe adequately. The leftmost panel of Figure 10 shows the observed score distribution for one data set along with the PPD of the score distribution. Figure 10 clearly indicates model misfit (and hence represents...
a success for the PPMC method), especially for the higher raw scores. There are more observed scores than expected near 20 and near 30 under the model and fewer observed scores than expected around 25. The middle panel of Figure 10 shows the standard $\mathcal{N}(0, 1)$ ability distribution on the same axes as the normal mixture. The unusual aspects of the observed score distribution relative to the 3PL model, assuming an $\mathcal{N}(0, 1)$ ability distribution, correspond directly to the areas in the ability distribution where the mixture differs from the typical distribution. The rightmost panel of Figure 10 shows the corresponding model check for the aggregate measure $\chi^2_{NC}$ that pools information across all score levels. The realized discrepancies are much larger than that predicted under the model. The pattern persists across data sets. The PPP-value for the aggregate measure $\chi^2_{NC}$ is less than 0.02 for all the 100 generated data sets.

**Real Data Example: Food Security Data**

The USDA’s U.S. Food Security Measurement Project administers a battery of survey questions as a supplement to the Current Population Survey (CPS) in an attempt to estimate the proportion of the population that is (in)secure in their ability to obtain “enough food for an active, healthy life.” Respondents are classified into one of three food security levels (food secure, food insecure without hunger, food insecure with hunger) based on their responses to the 18 items listed in Table 3.

The USDA uses the Rasch model to measure the unidimensional latent construct “food insecurity” by dichotomizing all of the items listed above. Responses of “often” and “sometimes” to Items 1 through 6 are treated as successes ($y_{ij} = 1$), and “never” is treated as a failure ($y_{ij} = 0$).
Table 3
The 18 Food Security Items

<table>
<thead>
<tr>
<th></th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>We worried whether our food would run out.</td>
</tr>
<tr>
<td>2</td>
<td>Food bought didn’t last.</td>
</tr>
<tr>
<td>3</td>
<td>Adult unable to eat balanced meals.</td>
</tr>
<tr>
<td>4</td>
<td>Relied on only a few kinds of low-cost foods to feed children.</td>
</tr>
<tr>
<td>5</td>
<td>Couldn’t feed child(ren) balanced meals.</td>
</tr>
<tr>
<td>6</td>
<td>Child(ren) not eating enough.</td>
</tr>
<tr>
<td>7</td>
<td>Adult cut size or skipped meals.</td>
</tr>
<tr>
<td>8</td>
<td>. . . this happened in three or more months over the past year.</td>
</tr>
<tr>
<td>9</td>
<td>Adult ate less than they felt they should.</td>
</tr>
<tr>
<td>10</td>
<td>Adult hungry but couldn’t afford to eat.</td>
</tr>
<tr>
<td>11</td>
<td>Adult lost weight because couldn’t afford to eat enough.</td>
</tr>
<tr>
<td>12</td>
<td>Adult didn’t eat for an entire day because couldn’t afford to eat.</td>
</tr>
<tr>
<td>13</td>
<td>. . . this happened in three or more months.</td>
</tr>
<tr>
<td>14</td>
<td>Cut the size of child(ren)’s meals.</td>
</tr>
<tr>
<td>15</td>
<td>Child(ren) was(were) hungry.</td>
</tr>
<tr>
<td>16</td>
<td>Child(ren) skipped meals.</td>
</tr>
<tr>
<td>17</td>
<td>. . . this happened in three or more months.</td>
</tr>
<tr>
<td>18</td>
<td>Child(ren) did not eat for entire day.</td>
</tr>
</tbody>
</table>

Note. Respondents were asked if Statements 1 through 6 were “often,” “sometimes,” or “never” true over the past 12 months. For the remaining items, except 8, 13, and 17, respondents responded by answering yes or no. Follow-up Items 8, 13, and 17 asked respondents whether the preceding statement was true in “only 1 or 2 months,” “some months but not every month,” or “almost every month.” Items 4 through 6 and 14 through 18 were presented only to those households in which there were children; the other 10 items were administered to all respondents.

Follow-up Questions 8, 13, and 17 are dichotomized by treating responses of “only in 1 or 2 months” as failures. The Food Security Project treats all missing responses as failures ($y_{ij} = 0$), and this study will do the same here.

A number of studies have suggested that the unidimensional Rasch model may not be adequate for the 18 questions listed above. Froelich (2002) investigated the unidimensionality assumption for the Food Security data from the 1999 CPS with the DETECT and DIMTEST (see, e.g., Stout et al., 1996, and the references therein) procedures and found evidence of at least two dimensions, with the items pertaining to children loading on one dimension and the items pertaining to adults/households loading on the second dimension. Nord (1999, as cited in Opsomer, Jensen, Nusser, Dringei, & Amemiya, 2002) noted that only individuals who pass a set of screening questions are administered the food security items. If it is believed that the entire population is normally distributed, then the mixing distribution employed in analysis should account for the nonrepresentative sample. Johnson (2004) employed a truncated normal distribution and fitted the 2PL model to Food Security data from households without children from the 2002 CPS. That paper found that differences in the discrimination parameters were statistically significant. Specifically, the paper found that the discrimination parameter for the item that asks if an “Adult was hungry but couldn’t afford to eat” was significantly larger than the discrimination parameters for all other nonchild food security items.

This section demonstrates how the biserial correlation, observed score distribution, and Mantel-Haenszel posterior predictive diagnostics are used to examine the appropriateness of the Rasch model for the Food Security data. The responses of 12,722 households (with children) from the 2003 CPS (available from http://www.census.gov/) are examined.

Figure 11, using the same symbols as in Figure 2, compares the observed biserial correlation with their posterior predictive distributions. It was noted in the discussion of simulation studies that the
biserial correlations for item responses generated from 2PL and 3PL items tend to be more variable than would be expected under the Rasch model. In the right panel of Figure 11, the observed standard deviation of the biserial correlations (designated with a vertical dashed line), although above the mean, is not terribly extreme when compared to the posterior predictive distribution of the same measure. However, the observed biserial correlations for all items except Item 4 lie above the 90% posterior predictive interval. In fact, the mean of the observed biserial correlations is larger than the mean of the generated biserial correlations in 100% of the posterior predictive samples. This may be a consequence of a nonnormal latent variable distribution as described below.

Figure 12, using the same symbols as in the leftmost plot in Figure 10, compares the observed raw score distribution with the posterior predictive score distributions. It is clear from the figure that the observed data set has significantly more respondents with a raw score of zero and significantly fewer respondents with scores 1 through 4 than expected under the analysis model. This finding is somewhat surprising because, under the Rasch model, the observed score is a sufficient statistic for the latent variable $\theta$, and under joint maximum likelihood estimation, the observed score distribution should be estimated perfectly. However, this analysis assumes a normal distribution for the latent variables, which, in this case, is not flexible enough to fit the observed raw score distribution.

Finally, the unidimensionality assumption is evaluated with the Mantel-Haenszel posterior predictive diagnostic plot in Figure 13. Both halves of the plot (above and below the diagonal) show the same PPP-values using circles as the half above the diagonal in Figure 9. The plot seems to indicate that there are some groups of items that are more closely related (empty circles) than expected under the unidimensional Rasch model that was fitted to the data. Specifically, Items 4 to 6 and Items 14 to 17 appear to form a single cluster of items that are more closely related than the Rasch model permits. These are all items that ask about the food availability for children in the home; Item 18 is the only child item that is not obviously part of the cluster. Items 7 through 13,
which ask specifically about the food intake of an adult in the household, appear to make up another item cluster; Item 3 is the only adult item not included in the cluster. The first two items may be viewed as a third cluster. These results are very similar to those found in Froelich (2002).

The posterior predictive diagnostics examined here present strong evidence against the use of a Rasch model with a normal latent variable distribution for the analysis of the Food Security data. However, it is not clear what extension of the model, if any, will correct the several problems uncovered by the posterior predictive diagnostics. One strategy might be to first fit a multidimensional Rasch model to address the findings of Figure 13, but it is not clear that a multidimensional model will correct the problems with the biserial correlations and the observed score distribution. Another strategy might use a more flexible distribution for the latent variable, one that can capture the shape of the observed score distribution. However, any such distribution will most likely fail to handle the problems identified in Figure 13. A combination of the two remedies might be needed. No matter what extension to the Rasch model is implemented, it remains important for the researcher to examine the appropriateness of the model assumptions.

**Summary and Recommendations**

This article, through several simulations and a real data example, demonstrates that PPMC provides a straightforward way to perform a collection of Bayesian model checks. The method is flexible enough to correctly detect lack of fit for a range of model violations that are of usual concern to psychometricians. One strong aspect of this article is the use of easily comprehensible and attractive graphical displays to present the model-checking results. Although the article deals with dichotomous items only, the techniques easily generalize to the case with polytomous items. Results in a companion article (Sinhaaray, 2005) applying the PP method to several real data examples provide further support for the findings from these simulations.
The choice of discrepancy measures is a vital issue in the application of the PPMC method. A number of discrepancy measures examined in this work are found useful in assessing IRT model fit. The OR and MH statistics, examining the association between item pairs, are found to successfully detect lack of fit whenever there is a violation of the LI assumption (e.g., for a multidimensional or a speeded test). The observed score distribution can be used to detect lack of fit when the assumed ability distribution is not correct. The discrepancy measures used in this work are natural quantities of interest in any educational assessment. Thus, another positive aspect of the model-checking approach presented here is that it provides a better understanding of the trade-offs among different types of IRT models. For example, the simulations suggest that when the items are of varying discrimination, then the Rasch model is not adequate to capture the dependencies in the data that are evident in the odds ratios. The simulations also show that the observed score distribution in the presence of guessing is not adequately described by the 2PL model, even though the model captures several other features of 3PL data.

The simulation results provide some advice concerning what has always been a difficult question about IRT model checking in real data settings—what aspects of model fit should be checked, and how should the checking be done? Naturally, the answer depends on an investigator’s circumstance. If there is a specific concern, then one ought to use discrepancy measures that address the concern. For example, researchers have questioned whether the unidimensional Rasch model is adequate for the Food Security data discussed in the preceding section, and the Mantel-Haenszel statistic appears to confirm that the data are not unidimensional. Another example is provided by a real data example in Sinharay (2005), where there were some questions about speededness of a 45-item basic skills test, and PPC using the OR statistic confirmed the problem.
If there are no specific a priori concerns for the model-checking exercise (as would happen with a model whose output is used for a number of different purposes), it is recommended to use a number of discrepancy measures (as suggested by van der Linden & Hambleton, 1997) for assessing different aspects of fit of the model; the simulations here suggest that the MH statistic and observed score distribution are especially useful. In the examination of the Food Security data, this study began by looking for evidence arguing against the form of the Rasch item response function but found instead, by examining the observed score distribution, that the assumption of a normal latent variable distribution is a bigger concern.

Only when the model fits all relevant aspects of the data adequately (as happened with a real example in Sinharay, 2005) can the investigator be confident that inferences made are reasonable. It is not entirely clear what to do when the model does not fit one or more aspects of the data. Possible actions include employing the model to make inferences but being aware of the limitations and reporting its deficiencies (Gelman et al., 1996, p. 799), discarding items or pooling testlet items with the hope that the chosen model will fit the revised data, or using a more general model such as the testlet model or multidimensional model. Another key issue when misfit is found is judging when discrepancies between the test data and the model fit are important enough to address. Such decisions require additional statistical analysis and subject matter expertise in addition to the model-checking plots and p values. Sinharay (2005) elaborates on this issue.

The results of this study should be of great interest to a range of researchers. Clearly, the results are of interest to researchers using a Bayesian approach. For any IRT model analyzed in a Bayesian way, the PPMC method is convenient because it builds on the posterior distribution obtained naturally during data analysis. With ever-increasing demand from test consumers, the need for more complicated models and hence the use of Bayesian methods is likely to accelerate; therefore, it will be to the advantage of the psychometric research community to develop model-checking strategies that take advantage of this. The findings of this study, especially about which diagnostics appear most helpful, should also be of interest to those using a frequentist approach (e.g., using maximum likelihood estimation). A variation of the PPMC method can be applied by using a point estimate for the model parameters (rather than using a sample from the posterior distribution of the parameters) to generate a number of simulated data sets. The simulated data sets then provide a reference distribution for model diagnostics. However, this frequentist approach, unlike the PPMC method, does not take into account uncertainty in the estimation of the parameters (e.g., Bayarri & Berger, 2000).

Notes

1. Those that are influenced by a single trait that is different from the trait that the test intends to measure; the latter trait is common to each item in the test.
2. Those that are influenced by different traits.
3. Both Clusters 1 through 15 and 16 through 30.

References


Chapman & Hall.


Acknowledgments

The authors thank Paul Holland, Shelby Haberman, Andreas Oranje, Ronald Hambleton, Hariharan Swaminathan, Charlie Lewis, Robert Mislevy, Roy Levy, Dan Eignor, Mark Reckase, and the three anonymous reviewers for their invaluable advice. The authors gratefully acknowledge the help of Deepayan Sarkar with the plots and Kim Fryer and Elizabeth Brophy with proofreading.

Author’s Address

Address correspondence to Sandip Sinharay, Educational Testing Service, MS-12T, Rosedale Road, Princeton, NJ 08541; e-mail: ssinharay@ets.org.