Simulation Studies
Applying Posterior Predictive Model Checking for Assessing Fit of the Common Item Response Theory Models

Sandip Sinharay
Matthew S. Johnson
Simulation Studies Applying Posterior Predictive Model Checking for Assessing Fit of the Common Item Response Theory Models

Sandip Sinharay
Educational Testing Service, Princeton, NJ

Matthew S. Johnson
Baruch College, New York, NY

October 2003
Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 7-R
Educational Testing Service
Princeton, NJ 08541
Abstract

Model checking in item response theory (IRT) is an underdeveloped area. No universal model-checking tool exists in this field. The posterior predictive model-checking method (Guttman, 1967; Rubin, 1981, 1984) is a popular Bayesian model-checking tool because of its simplicity, strong theoretical basis, obvious intuitive appeal, and ability to provide graphical evidence. This study applies the posterior predictive model-checking method to assess the fit of the popular IRT models (Rasch model, 2PL model, and 3PL model) for tests with dichotomous items. A series of simulation studies applying the method yields promising results. An important issue with the application of the posterior predictive model-checking method is the choice of discrepancy measures (which are like the test statistics in classical hypothesis testing). This article examines the performance of a number of discrepancy measures for assessing different aspects of fit of the common IRT models and suggests simple summarization of the results. Simple graphical summaries provide useful insight about the fit of the models. The odds ratios, corresponding to the responses of the examinees to pairs of items, appear to be powerful discrepancy measures in detecting several types of misfits of the IRT models. The observed score distribution and biserial correlation coefficients are also found useful as discrepancy measures on a number of occasions. A companion article applying the posterior predictive method to real data examples shows that the method has the potential to become very useful in psychometrics.

Key words: Bayesian methods, discrepancy measures, odds ratio, p-values, speededness, unidimensionality
Acknowledgements

A number of ideas in this work are outcomes of the first author’s conversations with Paul Holland. The authors thank Shelby Haberman, Andreas Oranje, Hal Stern, Hariharan Swaminathan, Charlie Lewis, and Dan Eignor for their invaluable advice.
1. Introduction

Model checking, or assessing the fit of a model, is a crucial part of any statistical analysis. Before making any conclusions from the application of a statistical model to a data set, an investigator should assess the fit of the model to make sure that the model can explain the different aspects of the data set adequately. Serious misfit (failure of the model to explain a number of aspects of the data that are of practical interest) should result in the replacement or extension of the model, if possible. Even if a model is the final model for an application, it is important to be aware of its limitations before making any inferences.

For statistical models used in psychometrics, and especially in item response theory (IRT), well-established statistical techniques for assessing model fit do not exist except for the simplest models (van der Linden & Hambleton, 1997, p. 16). The standard $\chi^2$ tests do not apply unless the number of items is small. Analysis of residuals have received occasional attention (van den Wollenberg, 1982; Molenaar, 1983; Reiser, 1996), but the applications have been limited. Therefore, there exists significant scope of further research in this area.

The posterior predictive model checking (PPMC) method is a popular Bayesian model checking tool because of its simplicity, strong theoretical basis, and close connection to the classical goodness-of-fit tests (Gelman, Meng, & Stern, 1996). The method primarily consists in comparing the observed data with replicated data, those predicted by the model, using a number of discrepancy measures; a discrepancy measure, like a classical test statistic, measures the difference between an aspect of the observed data set and a replicated data set. Practically, a number of replicated data sets are generated from the predictive distribution of replicated data conditional on the observed data (called the posterior predictive distribution). Any systematic differences between the observed data set and the replicated data sets indicate potential failure of the model to explain the data. Graphical display is the most natural and easily comprehensible way to examine the difference; another powerful tool is the posterior predictive $p$-value, the Bayesian counterpart of the classical $p$-value. Stern (2000) provides a list of examples in applied statistics where the PPCMC method has proved useful.

Discrepancy measures, the Bayesian counterpart of the classical test statistics, play
an important role in the PPMC method. In an application of the PPMC method, the key to success often is a set of discrepancy measures that effectively capture the aspects of the data that the model cannot explain satisfactorily. It is possible that a model does not adequately explain a data set, but some discrepancy measures will detect the misfit while some will not.

This work applies the PPMC method for assessing the fit of the most popular models (Rasch model, 2PL model, and 3PL model) in IRT through a series of simulation studies. A companion article (Sinharay & Johnson, in press) applies the method to real data examples. A major focus here is in finding a suite of discrepancy measures that will assess the aspects of model fit that the practitioners usually worry about in psychometrics, e.g., unidimensionality, local independence, speededness, etc. Posterior predictive checks result in many numbers and an important question is how to summarize the output in a precise and easily comprehensible manner. This work suggests attractive graphical approaches for summarizing the output. This report considers tests with dichotomous items only, although the concepts easily generalize to the case with polytomous items. Work about an item fit measure is in progress.

Important applications of the PPMC method in educational assessment include Rubin and Stern (1994), Hoijtink and Molenaar (1997), Scheines, Hoijtink, and Boomsma (1999), Albert and Ghosh (2000), Janssen, et al. (2000), Maris and Maris (2002), van Onna (2003), and Fox and Glas (2003), as the method gains increasing popularity in the field. However, none of the above works provides a comprehensive treatment of the PPMC method and all of them examine only a few discrepancy measures. This research takes a more general approach in attempting to provide researchers with a comprehensive toolkit for assessing fit of IRT models.

Section 2 describes the statistical models in item response theory that are of interest in this work. Section 3 introduces the posterior predictive model checking method and provides a detailed discussion of different aspects about the method. Section 4 discusses the discrepancy measures examined in this study. Section 5 describes an outline of the simulation studies aimed at finding powerful discrepancy measures for assessing different aspects of IRT model fit. Section 6 provides detailed description of the results from the
simulation studies. Section 7 provides the conclusions from the work and directions of future work.

2. Statistical Models in Item Response Theory

This section describes a general set up for a statistical model in IRT, thus defining the scope of this study. Consider an educational assessment consisting of \( J \) items that is given to \( I \) individuals. Individual \( i \) has a proficiency (ability) \( \theta_i \), possibly multidimensional, that the assessment attempts to measure. Item \( j \) has a corresponding item parameter vector \( \eta_j \), measuring aspects of the item like its difficulty, discrimination power, etc. For simplicity, this work assumes that every individual gets each item (this will not restrict the conclusions of this study in any way). Let \( y_{ij} \) denote the score of the \( i \)-th individual to the \( j \)-th item. This work deals with dichotomous scores only, i.e., the score is 1 if the examinee answered the item correctly and 0 otherwise.

The next step is to assume a probabilistic component \( P(y_{ij} | \theta_i, \eta_j) \) expressing the probability of a correct response of the \( i \)-th individual to the \( j \)-th item in terms of \( \theta_i \) and \( \eta_j \). For example, the 3-parameter logistic model (3PL; Lord, 1980) has three item parameters for each item, i.e., \( \eta_j = (a_j, b_j, c_j) \), \( a_j \), \( b_j \), and \( c_j \) respectively being the slope, difficulty and guessing parameters of item \( j \); assuming that the proficiencies \( \theta_i \)’s are unidimensional, the model expresses the probability of a correct response as

\[
P(y_{ij} \mid \theta_i, \eta_j) \equiv p(y_{ij} = 1 \mid \theta_i, \eta_j) = c_j + (1 - c_j) \logit^{-1}(a_j (\theta_i - b_j)),
\]

where \( \logit^{-1}(x) \equiv \exp(x)/(1 + \exp(x)) \).

Almost all of the popular models in IRT make the assumption of local independence, which means that given the proficiency variables, the observed outcomes for different items are independent. There are some models, however, that do not make this assumption, as we will see soon.

The Rasch model (Rasch, 1960), also referred to as the 1PL model, is a popular IRT model because of its simplicity, given by

\[
P(y_{ij} \mid \theta_i, \eta_j) = \logit^{-1}(a(\theta_i - b_j));
\]
The 2PL model (Birnbaum, 1968), given by
\[ P(y_{ij} \mid \theta_i, \eta_j) = \text{logit}^{-1}(a_j (\theta_i - b_j)), \]
is also extensively used in IRT.

The above three models are the most popular models in psychometrics and the testing companies around the world use these models operationally. One objective in this work will be to find out if these three models can perform adequately for real assessment data.

In this work, a standard normal population distribution assumption,
\[ \theta_i \sim \mathcal{N}(0,1), \]
fixes the scale of the problem for the 1-, 2-, and 3PL model. In other words, this work fits the model (as in Holland, 1990b)
\[ \int \prod_{j=1}^{J} P(y_{ij} \mid \theta_i, \eta_j)^{y_{ij}} (1 - P(y_{ij} \mid \theta_i, \eta_j))^{1-y_{ij}} \phi(\theta) d\theta \]
to the response vector of examinee \( i, i = 1, \ldots I \), treating the examinee proficiencies as nuisance parameters, where \( \phi(x) \) denotes the ordinate of the \( \mathcal{N}(0,1) \) distribution at \( x \).

A testlet originally meant a group of items on a content area that is developed as a unit and a fixed number of predetermined paths that an examinee may follow (Wainer & Kiely, 1987). Bradlow, Wainer, and Wang (1999) use the term testlet to mean a subset of items generated from a single stimulus, e.g., a reading comprehension passage. This work will use the latter definition henceforth.

Bradlow et al. (1999) argue that the assumption of local independence is violated for assessments involving testlets; they suggest the testlet model to take into account the dependence of the responses of an examinee to the items within a testlet. Suppose the item \( j \) belongs to the testlet \( d(j) \). The testlet model expresses the probability of success of an individual on an item as
\[ P(y_{ij} = 1 \mid a_j, b_j, c_j, \theta_i, \gamma_{id(j)}) = c_j + (1 - c_j) \text{logit}^{-1}\{a_j (\theta_i - b_j - \gamma_{id(j)})\} \tag{4} \]
The term \( \gamma_{id(j)} \) is the testlet effect (interaction) of testlet \( d(j) \) with person \( i \). Further, the testlet model assumes the population distributions \( \theta_i \sim \mathcal{N}(0,1) \) and \( \gamma_{id(j)} \sim \mathcal{N}\left(0, \sigma^2_{d(j)}\right) \).
This work also considers a multidimensional extension of the 3PL model, the linear logistic 2-dimensional model (Reckase, 1997, and the references therein), given by

\[
P(y_{ij} \mid \theta_i, \eta_j) = c_j + (1 - c_j) \logit^{-1} (a_{1j}\theta_{1i} + a_{2j}\theta_{2i} - b_j),
\]

\[
\eta_j = (a_{1j}, a_{2j}, b_j, c_j); \quad \theta_i = (\theta_{1i}, \theta_{2i}) \sim \mathcal{N}_2(0, 0, 1, 1, \rho), \quad -1 \leq \rho \leq 1.
\]

3. Posterior Predictive Model Checking Techniques

Guttman (1967) first suggested the idea of the posterior predictive distribution; he used the terminology density of a future observation to describe the concept. Rubin (1981) applied the idea of the posterior predictive distribution to formulate the posterior predictive model checking (PPMC) method and gave a formal Bayesian definition of the technique (1984). In the latter work, he argues that the PPC method is Bayesianly justifiable, although not in the usual sense, and is Bayesianly relevant. Gelman et al. (1996) extended the PPC method to allow for more direct assessment of the discrepancy between the data and the posited model.

In Bayesian statistics, a researcher can check the fit of the model in one of three ways (Gelman et al., 1996): (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model can explain the data adequately. PPC techniques address the third of these concerns.

The Description of the PPC Method and the Posterior Predictive P-values

Let \( p(y \mid \omega) \) denote the likelihood distribution for a statistical model applied to data (examinee responses) \( y \), where \( \omega \) denote all the parameters in the model. Let \( p(\omega) \) be the prior distribution on the parameters. Let \( p(\omega \mid y) \equiv \frac{p(y \mid \omega)p(\omega)}{\int_{\omega} p(y \mid \omega)p(\omega) \, d\omega} \) denote the posterior distribution of \( \omega \). Let \( y^{rep} \) denote replicated/future data that one might observe if the process that generated the data \( y \) is replicated.

The PPC method suggests checking Bayesian models using the posterior predictive distribution (or the predictive distribution of replicated data conditional on the
observed data)

\[
p(y^\text{rep}|y) = \int p(y^\text{rep}, \omega|y) d\omega = \int p(y^\text{rep}|\omega, y)p(\omega|y) d\omega = \int p(y^\text{rep}|\omega)p(\omega|y) d\omega \quad (6)
\]
as a reference distribution for the observed data \(y\).

The replication \(y^\text{rep}\) here represents replicate data that we might observe if the experiment that generated \(y\) is replicated with the same value of \(\omega\) that generated the observed data. Since the value \(\omega\) that generated the observed data is unknown, the PPMC method derives the posterior (given \(y\)) predictive distribution of \(y^\text{rep}\) by averaging the likelihood over the plausible values of \(\omega\), where the plausible values are quantified by the posterior distribution \(p(\omega|y)\).

The next step in the PPMC method is to compare the observed data \(y\) to its reference distribution. In practice, as in the prior predictive method, test quantities or discrepancy measures \(D(y)\) are defined and the observed value of \(D(y)\) is compared to the reference distribution of \(D(y^\text{rep})\), any significant difference between them indicating a model failure.

Because of the difficulty in dealing with (6) analytically for all but simple problems, Rubin (1984) suggests simulating replicate data sets from the posterior predictive distribution in practical applications. One draws \(N\) simulations \(\omega^1, \omega^2, \ldots, \omega^N\) from the posterior distribution \(p(\omega | y)\) of \(\omega\), and then draws one \(y^\text{rep}\) from the predictive distribution \(p(y | \omega)\) using each simulated \(\omega\). The process results in \(N\) draws from the joint posterior distribution \(p(y^\text{rep}, \omega | y)\), or, equivalently, from \(p(y^\text{rep}|y)\). The posterior predictive check boils down to comparing the values of the observed discrepancy \(D(y)\) and the replicated discrepancy measures \(D(y^\text{rep,n})\), \(y^\text{rep,n}\), denoting the replicated data set generated on the \(n\)-th draw, \(n = 1, 2, \ldots, N\). Any significant difference between the observed discrepancy and the replicated discrepancies indicates a possible failure of the model.

From the above description, it is clear that the PPMC method combines very well with the Markov chain Monte Carlo (MCMC) algorithms (Gilks, Richardson, & Spiegelhalter, 1996; Gelman, Carlin, Stern, & Rubin, 1995). An MCMC algorithm generates draws from the posterior distribution of \(\omega\), which is the first step in the PPMC method. The next step in the method is to generate, for each iteration of the MCMC, a replicated
data set $y^{rep}$ using the generated parameter value in that iteration. The last step is to compare the replicated data sets to the observed data set.

One way to perform the comparison is to plot the generated values of $D(y^{rep,n})$ in a histogram and examine where $D(y)$ lies with respect to the histogram—an inadequate model and a powerful discrepancy measure will result in $D(y)$ lying at the tail area of the histogram. Gelman et al. (1996) suggest that a graphical comparison should be done wherever it is possible. Stern (2000) comments that the PPMC method is in the spirit of a diagnostic plot rather than a test.

Another popular summary of the comparison is the tail-area probability or posterior predictive p-value (PPP-value), the Bayesian counterpart of the classical p-value (Bayesian p-value):

$$p_b = P(D(y^{rep}) \geq D(y)|y) = \int \int I_{D(y^{rep}) \geq D(y)} p(y^{rep}|y) dy^{rep},$$

where $I_{D(y^{rep}) \geq D(y)}$ denotes the indicator function for the event A. The above definition suggests that the posterior predictive p-value is also the expectation of $I_{D(y^{rep}) \geq D(y)}$, where the expectation is with respect to the posterior predictive distribution $p(y^{rep}|y)$. As an immediate consequence, if an investigator has generated N replicate data sets from the posterior predictive distribution, the proportion of the N replications for which $D(y^{rep,n})$ exceeds $D(y)$ provides an estimate of the PPP-value. Very extreme posterior predictive p-values (close to 0 or 1) indicate model misfit.

Gelman et al. (1996) extend the posterior predictive approach to use realized discrepancies $D(y, \omega)$ that depend on the data and the parameters, but this paper does not consider any realized discrepancies.

**A Simple Example From Meng (1994)**

Let $y_1, y_2, \ldots, y_n$ denote a number of independent and identical observations from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$. In the context of psychometrics, the $y_i$s may denote the estimated coaching effects (for an assessment) of a number of coaching centers. Suppose the goal is to detect if there is enough evidence to
effect is larger than 0.

Hence, there is evidence that the average value of the sample mean using a vertical line. The sample mean indeed looks extreme with

Example—Posterior predictive density and observed value of sample mean.

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

When the posterior predictive density of the sample mean along with the observed

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

strategically way to perform the task. For example, if \( \theta = \mu \), a strategy above. Furthermore, the posterior predictive density of

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

is a scaled t-distribution with \( n \) degrees of freedom. The posterior predictive distribution of \( \theta \), obtained by integrating out \( \theta \) with

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

the posterior sample mean, is defined as a t-distribution under the null

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

where \( \mathbf{y} \) is the observed data. The posterior predictive check then amounts to examining if the observed

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

is given by

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

the mean here is the sample mean, i.e., the noninformative prior distribution on \( \theta \). A possible test statistic of discrepancy

from the viewpoint of a frequentist, from the viewpoint of an investigator applying

\[ p(y \mid \theta, \mathbf{y}) \propto p(y \mid \theta) p(\theta \mid y) \]

conclude that the average coaching effect is larger than 0 (i.e., \( H_0 \): \( \theta = 0 \) and \( 0 < \theta \) : \( \theta \)).
The posterior predictive p-value for the example, using (7) and (8), is

\[
P(\bar{y}^{rep} > \bar{y} \mid y) = 1 - \Upsilon_n \left( \frac{\sqrt{n} \bar{y}}{\sqrt{\sum y_i^2 / n}} \right),
\]
where \(\Upsilon_n(\cdot)\) denotes the cumulative distribution function of the t distribution with \(n\) degrees of freedom.

Remember that the classical test statistic for this problem is the pivotal (i.e., one whose distribution does not depend on the parameters of the model) t-statistic \(T = \frac{\bar{y}}{s} \sqrt{n}\), where \(s\) is the sample standard deviation with divisor \((n-1)\). Under the null hypothesis, \(T \sim t_{n-1}\) and hence the classical p-value is

\[
P \left( T > \frac{\bar{y}}{s} \sqrt{n} \right) = 1 - \Upsilon_{n-1} \left( \frac{\sqrt{n} \bar{y}}{s} \right).
\]

The difference in (9) and (10) maybe significant for small sample size; especially, they may lead to different conclusions for the same data set. For example, if \(n = 4, \bar{y} = 2, s = 1\), then the PPP-value is 0.07 while the classical p-value is 0.01. However, for even a moderately large sample size, the difference is inconsequential. For example, if \(n = 10, \bar{y} = 2, s = 1\), (which are the values corresponding to Figure 1) then the PPP-value is 0.009 while the classical p-value is 0.00005; i.e., the PPP-value is larger than the classical p-value, but both of them provide strong evidence against the null hypothesis.

However, consider the discrepancy measure \(D(y) = \frac{\bar{y}}{s}\), scaled and hence more reasonable, for use with the PPMC method. Because \(\frac{\bar{y}}{s} \sqrt{n} \sim t_{n-1}\),

\[
p \left( \sqrt{n}D(y^{rep}) \mid y \right) = p \left( \sqrt{n}D(y^{rep}) \right) \equiv t_{n-1},
\]
the same distribution as that of the classical test statistic, and the PPP-value becomes the same as (10), the classical p-value.

**Properties of PPP-values and Their Relationship With Classical P-values**

It is straightforward to observe that if the discrepancy measure is a pivotal quantity, then the PPP-value is identical with the classical p-value (as the simple example earlier shows). Meng (1994) provides a comparison of the classical p-values and the Bayesian p-values. He argues that the PPP-value can also be viewed as the posterior mean of a
classical p-value, averaging over the posterior distribution of (nuisance) parameters under
the null model. He also performs frequency evaluations to show that in general, under
certain assumptions, the Type I frequentist error of an $\alpha$-level posterior predictive check is
often close to but less than $\alpha$ and will never exceed $2\alpha$. Gelman et al. (1996) comment,
mainly on the basis of Meng (1994), that “empirical and theoretical studies so far do
suggest that PPP-values generally have reasonable long-run frequentist properties.”

Gelman et al. (1996, p. 174) suggest that the posterior predictive checks are
generalizations of classical tests in that they average over the posterior distribution of
the unknown parameter rather than fixing it at some value (like the MLE, as is the
practice with classical testing). Gelman et al. (1996) argue that the posterior predictive
replications appear to be the replications that the classical approach intends to address.
Posterior predictive checks do not depend on the clever construction of pivotal quantities
or on asymptotic results like in classical testing, and are therefore easily applicable to any
probability model, however complicated the model is.

However, posterior predictive checks have their own problems. The checks are
time-consuming. Also, the choice of the discrepancy measures and appropriate predictive
distributions requires careful consideration of the type of inferences in which the researcher
is interested.

*Effect of Prior Distributions*

Because the posterior predictive checks are Bayesian methods by nature, a question
arises about the sensitivity of the results obtained to the prior distributions on the model
parameters. Gelman et al. (1996) comment that if the parameters are well-estimated
from the data, posterior predictive checks give results similar to classical model checking
procedures for reasonable prior distributions. Strongly informative prior distributions may
seriously affect the results of posterior predictive checks. The replicated data sets obtained
under strong incorrect prior distributions maybe quite far from the observed data. Gelman
et al. (1996, p. 757) provide such an example. On the contrary, a strong prior distribution,
if trustworthy, can help a researcher assess the fit of the likelihood part of the data more
effectively (Gelman et al., 1996).
For the IRT models, there exists a large amount of prior knowledge from previous administrations of different tests and this work uses that to form prior distributions for the parameters. For example, it is well-known in psychometrics that the guessing parameter (for the 3PL model) of a 5-option multiple choice item is expected to be around 0.2—we make use of that fact to form the respective prior distribution. Also, as the number of examinees is large for almost all the analyses in this work, the prior distributions have little effect on the posterior distributions, and hence on the posterior predictive data sets. Therefore, the prior distributions are not of much concern in this work.

Discussion

Meng (1994) suggests that uniformity under the correct model assumption (i.e., when “the null hypothesis $H_0$ is true” from a frequentist’s viewpoint) is a useful criterion for a $p$-value. He and Robins, van der Vaart, and Ventura (2000) show that the posterior predictive $p$-values (PPP-value) are not always uniformly distributed under the null hypothesis, not even asymptotically, and hence do not possess a desirable property of a $p$-value. The authors also suggest proper centering of the discrepancy measures to make them asymptotically uniform under the null model, although the revised measures are computationally prohibitive. Several authors, e.g., Bayarri and Berger (2000) and Robins et al. (2000) criticized the PPMC method for being conservative and using the data twice. Robins et al. (2000) also show that the PPP-values are conservative (i.e., often fails to detect model misfit), even asymptotically, for some choices of discrepancy measures—for example: (i) when the asymptotic mean of the discrepancy depends on the model parameters, and (ii) when the discrepancy and maximum likelihood estimator of the parameter are correlated. Bayarri and Berger (2000) propose alternative model checking strategies that are less conservative and produce $p$-values having asymptotic uniform distribution under the null hypothesis. Their approaches however are more difficult to carry out and interpret.

Even with the above limitations, the simplicity, the ease of computation, easy interpretability and ability to provide graphical evidence (as well as $p$-values) for suggesting limitations of a hypothesized model make the PPMC method a natural choice (Stern, 2000).
Also, there is nothing inappropriate in the double use of the data in the PPMC method. For some real applications, specially those using MCMC algorithms to fit rather complicated models (often hierarchical), the PPMC method is the only tool available. Also, it is possible to examine any function of data and parameter using the method. The nonuniformity of the PPP-value under the null can make the calibration difficult in borderline cases, but because it is a valid probability of an event, the PPP-value remains a fairly natural concept for data analysts to consider. The PPMC method is especially useful if one thinks of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

**Discrepancy Measures**

Technically, any function of the data and the parameters can play the role of a discrepancy measure in posterior predictive checks. However, choice of discrepancy measures is very important with the PPMC method. Virtually, all models are wrong and a statistical model applied to a data set usually explains certain aspects of the data adequately and some others inadequately. The challenge to the researcher applying the PPMC method to a problem is to find out discrepancy measures that have the power to detect the aspects of the data that the model cannot explain satisfactorily. Discrepancy measures that relate to features of the data not directly addressed by the probability model are expected to perform better (Gelman et al., 1995). For example, the percentage-correct scores for the items for the 1-, 2-, and 3PL models will be weak discrepancy measures—the difficulty parameters in these models address the percentage-correct scores, and the models will always reproduce the latter quantities. Failure to find the proper discrepancy (and number of applications of other discrepancies that show no sign of misfit) may lead to the incorrect conclusion that the model fits the data adequately.

For a practical problem, a good strategy is to examine a number of discrepancies corresponding to aspects of practical interest as well as some standard checks on overall fitness (Gelman et al., 1996). For example, if a researcher is worried that the unidimensionality assumption may not hold for a particular problem, it will be wise to examine discrepancies, if possible, that can detect the lack of unidimensionality. If those
discrepancies and a few others examining overall fitness of the model provides little evidence against the fit of the model, there is every reason to be confident that the model explains the data adequately for the purpose of interest.

For the popular statistical models in IRT, a number of discrepancies may be of interest, depending on the context of the problem. Interest may be in testing for one or more of overall model fit, item fit, unidimensionality, local independence, speededness, differential item functioning, person fit, etc., and different discrepancies may be effective for examining the different types of fit. Also, it is possible to employ overall discrepancies or those based on items, individuals, and interactions between them. One objective of this paper is to find discrepancy measures that are effective to assess the above-mentioned aspects of IRT model fit.

4. Discrepancy Measures Examined

This work examines a number of discrepancy measures, all of which are based on data \( y \) only (and no parameters). This section describes in detail the discrepancies and discusses the situations where they might be powerful.

_Percentage-correct Score for the Items_

Any reasonable model applied to an assessment data should predict the total scores (i.e., the number of examinees getting the item correct), or, equivalently, the percentage-correct scores for the items reasonably well. The total scores for the items are sufficient statistics for the difficulty parameters in the simplest possible IRT model, i.e., the Rasch model given in (2). Because almost all parametric IRT models build on the Rasch model, the model-predicted percentage-correct scores are expected to match the observed percentage-correct scores, especially because each item is usually given to a large number of individuals in any real-life educational assessment. Therefore, this measure will probably not be very powerful. Practically, examining this measure may help check the computations; if computations show that the model cannot predict the percentage-correct scores for the items well, that will be a strong indication of some problems in the computer
program used.

**Observed Score Distributions**

The observed score distribution (Lord, 1980), i.e., the distribution of the number of examinees obtaining exactly a certain number of items correct, is a natural quantity to use as a discrepancy measure. Consider a $J$-item test with each item, dichotomous, given to $I$ examinees. Denote $NC_j$ to be the number of examinees getting exactly $j$ items correct. Suppose $\text{NC} = (NC_0, NC_1, \ldots NC_J)'$. A reasonable model should predict the vector $\text{NC}$ well. Cressie and Holland (1983) prove $\text{NC}$, together with the percentage-correct scores for the items, to be jointly sufficient for the parameters of a Rasch model (the simplest IRT model practically used). Therefore, it is possible that any IRT model, even one not adequate for the data at hand, will predict $\text{NC}$ well and these measures will not be powerful.

**Biserial Correlation Coefficient**

In classical item analysis—a simple quantity that helps test developers judge how closely performance on a test item is related to the raw score—is the Pearson product moment correlation coefficient between the item score and the total score or the point biserial correlation, denoted as

$$r_{pbs} = \frac{(av_+ - av)}{sd} \sqrt{\frac{\hat{p}}{1 - \hat{p}}},$$

where $av_+$ is the average raw score for the examinees who answer the item correctly, $av$ is the average raw score of all of them, $sd$ is their standard deviation, $\hat{p}$ is the observed proportion-correct for the item. The point-biserial correlation of an item then is an indicator of its discrimination power—a good item will have high $r_{pbs}$. For tests with small number of items (perhaps 25 or less), Crocker and Algina (1986, p. 317) recommend examining the correlation between the item score and the total score with that item removed.

Assuming that the latent variable underlying item performance is normally distributed, Pearson (1909) derives a formula for the correlation between the latent variable
and a continuously distributed criterion such as a test score. The quantity, called the 
*biserial correlation coefficient*, is computed as:

\[ r_{bis} = r_{bis} \times \frac{\sqrt{\hat{p}(1 - \hat{p})}}{\phi(\Phi^{-1}(q))}, \]

where \( \Phi^{-1}(q) \) is the \( q \)-th quantile of the standard normal distribution.

The biserial correlation coefficients will obviously be good discrepancy measures to
detect the fit of a Rasch model when the items actually have varying discrimination powers,
because the Rasch model assumes all the items to have the same discrimination power. In
that situation, the Rasch model will severely underpredict quantities like the variance of
the biserial correlation coefficients. Albert and Ghosh (2000) find the variation of the point
biserial correlations to be a powerful measure for detecting misfit of the Rasch model. For
models with a discrimination parameter for each item (e.g., the 2PL model and the 3PL
model), this measure may not provide much insight about any misfit.

Other than the biserial correlation coefficients (both including the item and
excluding it) themselves, this study also examines their mean, variance, the minimum
and maximum of them (one-number summaries of the biserial correlation coefficients) as
discrepancy measures.

*Interaction/Association Among the Items*

There are no parameters in any IRT model to directly address how items
interact/associate with each other. Therefore, according to an earlier argument, any
discrepancy measure that captures the associations (interaction effects) among the items
will probably be effective in detecting possible misfit of an IRT model. This work examines
two such discrepancy measures, both described below.

*Proportion of Individuals Answering Pairs of Items Correctly*

van den Wollenberg (1982) and Molenaar (1983) examine classical test statistics
based on the number of individuals answering pairs of items correctly conditional on their
raw scores for detecting possible misfit of Rasch models. They find the statistics to be
sensitive to violation of local independence and unidimensionality for Rasch models. Reiser
(1996) suggests analysis of residuals based on number of individuals answering pairs of items correctly (referring to them as second-order marginals) for item response models in general. This work examines the effectiveness of the second-order marginals as discrepancy measures with posterior predictive checks. These discrepancies may be successful if the model cannot adequately explain the association among the items in a test, for example, when a univariate model is applied to a test assessing multiple skills. In that case, the actual number of examinees answering pairs of items that measure the same skill will probably be larger than what a univariate model can predict.

**Odds Ratio for Measuring Association Among Item-pairs**

Consider the item pair consisting of items \(i\) and \(j\) in a test. Denote \(n_{kk'}\) to be the number of individuals getting a score of \(k\) on item \(i\) and \(k'\) on item \(j\), \(k, k' = 0, 1\). We define

\[
OR_{ij} = \frac{n_{11}n_{00}}{n_{10}n_{01}}.
\]

The quantity on the right-hand side in the above definition is a sample odds ratio (see, for example, Agresti, 2002, pp. 45) corresponding to the population odds ratio:

\[
\frac{P(\text{item } i \text{ correct }|\text{ item } j \text{ correct})}{P(\text{item } i \text{ wrong }|\text{ item } j \text{ correct})} \cdot \frac{P(\text{item } i \text{ correct }|\text{ item } j \text{ wrong})}{P(\text{item } i \text{ wrong }|\text{ item } j \text{ wrong})}.
\]

This work investigates the performance of the sample odds ratio (referred to as the odds ratio hereafter) as a discrepancy measure with posterior predictive checking. Odds ratio is a measure of association—therefore, examining it should help a researcher to detect if the fitted model can adequately explain the association among the test items. For example, if a test measures multiple skills, a unidimensional model fitted to the test data will probably underpredict the odds ratios within groups of items testing each of the dimensions.

This work also examines an additive version of the odds ratio, i.e., the quantity

\[
n_{11} + n_{00} - n_{10} - n_{01}.
\]

This measure should provide the same inference as the odds ratios.

Also, the odds ratio (its additive version) is in some sense scaled (centered) version of the corresponding second-order marginal. Therefore remembering the simple example
earlier (where the scaled version of the sample mean was more powerful than the sample mean itself) and arguments from Robins et al. (2000), there is a possibility that the odds ratio (and its additive version) will be more powerful than the second-order marginal as a discrepancy measure.

5. Outline of the Simulation Studies

Before applying the discrepancy measures to real data sets, it is important to have an idea about the level and power of them, especially keeping in mind the fact that the PPP-values are usually not uniformly distributed under the null model and are conservative. Hence, this work performs detailed simulations to study the performance of the different discrepancy measures for a number of combinations of a data-generating model and an analysis model. There is a considerable difference of these simulation studies with many simulation studies in psychometrics literature regarding model checking. The latter studies attempt to find the unknown null distribution of the test statistics through the studies; in contrast, the PPP-values are valid probability statements and there is no need to find their distributional properties. The simulation studies here aim at examining the power or effectiveness of the discrepancy measures under different situations in an attempt to explore suitable discrepancy measures for assessing different aspects of IRT model fit.

Consider a data generating model $M_g$ and an analysis model $M_a$, where each model may be one among those discussed in Section 2. The following steps describe the details of the simulation study for the model pair:

1. A C++ program generates 100 data sets 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 range and combination of the item parameters of $M_g$ 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 from the National Assessment in Educational Progress, an ongoing educational survey administered by the National Center for Education Statistics (NCES), as the generating parameter values. A rich variety of generating parameters values ensure that
the findings of the simulation study are not limited and compensates the fact that the 
program uses only one set of generating parameter values to generate all 100 data sets 
for any pair \( \{M_g, M_a\} \).

2. For each of the above 100 data sets generated, the above-mentioned program computes 
the values of observed discrepancy measures and then fits the model \( M_a \) using an 
MCMC algorithm; the program runs five chains of length 4,000, discards the first 2,000 
draws in each chain and then thins the remaining of the output by retaining every 10th 
draw to get a combined posterior sample of size 1,000. A number of MCMC convergence 
diagnostics (Time-series plots, Gelman-Rubin convergence diagnostics, and Brooks-
Gelman multivariate potential scale reduction factor; see, for example, Sinharay, 2003, 
and the references therein) indicate that the above numbers are sufficient to ensure 
convergence for this example. There is a slight Monte Carlo error because of using a 
sample size of 1,000, but an increase in the size of the final posterior sample hardly 
results in any changes in the posterior predictive p-values. That is partially because 
there is effectively no difference between a value of 0.01 and 0.03 because both of them 
indicate significant problems with the model.

The prior distributions used are:

\[
\log(a_j) \sim \mathcal{N}(0, 1), \quad b_j \sim \mathcal{N}(0, 1), \quad \logit(c_j) \sim \mathcal{N}(-1.39, 1).
\]

The prior distributions are chosen from our experience of analyzing data from multiple 
choice tests. In the prior distribution for \( \gamma \), the mean value of -1.39 is an outcome of the 
fact that in multiple choice items with five options, the guessing parameter is expected 
to be around 0.2 and that \( \logit(0.2) = -1.39 \). Because the number of examinees is 
as large as 2,500 in these simulation studies, the prior distributions hardly affect the 
results of the posterior predictive checks.

3. For each of the draws in the final posterior sample, the program generates a replicated 
data set and computes values of the different discrepancy measures. The program then 
computes the posterior predictive p-values (PPP-value) for the discrepancies. The 
process results in a PPP-value for each discrepancy measure (computed using the final
A set of programs written with the open-source statistical package R (http://cran.r-project.org) summarize, both graphically and numerically, the output of the C++ program to provide useful feedback about the performance of each discrepancy measure. For each discrepancy measure, it is possible to plot the 100 PPP-values (obtained from the 100 generated data sets) using a histogram-type plot; alternatively, the mean, quantiles, and the proportion of times they are extreme (thus indicating a problem with the model) are useful summaries.

This study performs the above process for a number of combinations of $M_g$ and $M_a$. The whole process for a specific combination of data model and analysis model takes about 4 days on a Sun Blade 1000 workstation equipped with 950 MHz CPU and half gigabyte of RAM. The combinations in Table 1 with an asterisk ("*") in the corresponding cells are the cases examined. For example, we examine the case when the data-generating model is the 3PL model and the analysis model is the Rasch model; but we do not examine the case when the data-generating model is the testlet model and the analysis model is the Rasch model. The analysis model is always one among the 1PL, 2PL, and 3PL models, because our focus is to find out if these simpler models can adequately explain different types of test data; fitting any one among the other models is tedious. Also, it is intuitively clear that a 2PL model will explain data generated from an 1PL model (as the latter is a special case of the former)—so we do not examine that case. The same is true for other similar cases.

Table 2 shows the generating parameter values when the generating model is a 3PL model. To generate data from a 1PL or 2PL model, the data generator uses the relevant
column(s) of values of the parameters from Table 2. For example, the column of values of $b_j$ in the table are the generating parameters for data from a Rasch (1PL) model. The ability parameters are generated from a $N(0, 1)$ distribution.

**Table 2: The Generating Parameter Values for the Simulation Studies**

<table>
<thead>
<tr>
<th>Item ID</th>
<th>$a_j$</th>
<th>$b_j$</th>
<th>$c_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.70</td>
<td>1.81</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>1.89</td>
<td>-0.52</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>1.32</td>
<td>0.26</td>
<td>0.14</td>
</tr>
<tr>
<td>4</td>
<td>1.36</td>
<td>-1.48</td>
<td>0.17</td>
</tr>
<tr>
<td>5</td>
<td>1.17</td>
<td>-0.52</td>
<td>0.17</td>
</tr>
<tr>
<td>6</td>
<td>0.56</td>
<td>0.44</td>
<td>0.14</td>
</tr>
<tr>
<td>7</td>
<td>1.10</td>
<td>-2.15</td>
<td>0.30</td>
</tr>
<tr>
<td>8</td>
<td>1.68</td>
<td>0.96</td>
<td>0.25</td>
</tr>
<tr>
<td>9</td>
<td>1.01</td>
<td>-0.87</td>
<td>0.10</td>
</tr>
<tr>
<td>10</td>
<td>0.88</td>
<td>1.70</td>
<td>0.22</td>
</tr>
<tr>
<td>11</td>
<td>1.19</td>
<td>-2.41</td>
<td>0.23</td>
</tr>
<tr>
<td>12</td>
<td>0.60</td>
<td>-0.56</td>
<td>0.21</td>
</tr>
<tr>
<td>13</td>
<td>1.49</td>
<td>-0.27</td>
<td>0.12</td>
</tr>
<tr>
<td>14</td>
<td>2.01</td>
<td>0.04</td>
<td>0.20</td>
</tr>
<tr>
<td>15</td>
<td>2.40</td>
<td>-0.79</td>
<td>0.18</td>
</tr>
<tr>
<td>16</td>
<td>2.00</td>
<td>-0.38</td>
<td>0.17</td>
</tr>
<tr>
<td>17</td>
<td>1.48</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>18</td>
<td>1.10</td>
<td>-1.53</td>
<td>0.14</td>
</tr>
<tr>
<td>19</td>
<td>1.52</td>
<td>0.11</td>
<td>0.20</td>
</tr>
<tr>
<td>20</td>
<td>1.45</td>
<td>-0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>21</td>
<td>0.80</td>
<td>-0.52</td>
<td>0.11</td>
</tr>
<tr>
<td>22</td>
<td>0.67</td>
<td>-0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>23</td>
<td>0.83</td>
<td>-1.25</td>
<td>0.18</td>
</tr>
<tr>
<td>24</td>
<td>1.17</td>
<td>-0.52</td>
<td>0.13</td>
</tr>
<tr>
<td>25</td>
<td>1.43</td>
<td>-0.27</td>
<td>0.15</td>
</tr>
<tr>
<td>26</td>
<td>2.4</td>
<td>-0.44</td>
<td>0.40</td>
</tr>
<tr>
<td>27</td>
<td>1.53</td>
<td>1.75</td>
<td>0.25</td>
</tr>
<tr>
<td>28</td>
<td>1.2</td>
<td>-0.80</td>
<td>0.24</td>
</tr>
<tr>
<td>29</td>
<td>1.7</td>
<td>0.40</td>
<td>0.16</td>
</tr>
<tr>
<td>30</td>
<td>2.05</td>
<td>-0.93</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Figure 2 shows the corresponding item characteristic curves (ICCs). The figure plots all the true ICCs for the Rasch model together and then does the same for the 2PL model and the 3PL model, to provide an idea about the variation in the item parameters used in the simulation study.
measure are consistently extreme for any of the three situations (i.e., under the "true" model view) using the discrepancy measures. If the PFA-values correspond to a discrepancy will wrongly detect misfit of the model (i.e., "common a type I error" from a frequentist framework) and provide feedback regarding how often an investigator will conduct such investigations in the future.

**Case: Data Model 1PL/P2PL/3PL, Analyses Model Same as Data Model**

The odds ratio, for the items, (b) proportion correct scores for item pairs, and (c) additive version of the odds ratio does not show any results for the three measures (a) proportional correct scores. Therefore, the additive version of the odds ratio has considerable power, but the odds ratio measure (second-order marginals) is not very effective in assessing the independence of a model.

The proportion correct scores of all combinations of models examined here, the proportion correct scores for item pairs, and the percentage-correct scores for the items are reproduced very well. The discrepancy measures found powerful for the combination of the data model and analyses.

**Results from the Simulation Studies**

**Figure 2:** True ICCs of all the items in the simulation study.
assumption), it will be obvious that the measure has high Type I error associated with it and hence is inappropriate.

However, these analyses show that the discrepancy measures examined do not exhibit such behavior at all. The median of the PPP-values are close to 0.5 for almost all the measures and the proportion of times they are extreme (more than 0.95 or less that 0.05) are almost always less than 0.10. These results ensure that the Type I error rate is at the intended level for these discrepancy measures.

As an example, Figure 3 shows a histogram plotting the 100 PPP-values for the odds ratio for item pair \{1, 6\} when both the data model and the analysis model are the 3PL model.

![Histogram](image)

**Figure 3:** Example of p-values being spread uniformly over (0,1).

Additionally, for most of the measures, the PPP-values are more or less uniformly distributed between 0 and 1. There are some exceptions. For example, the PPP-values for the percentage-correct scores for the items (not a good choice as a discrepancy measure, as argued earlier) are more closely concentrated than a Uniform(0,1) random variable. Figure 4 shows the 100 PPP-values for this discrepancy measure for Item 5 when both the data model and the analysis model is the 3PL model—the values are tightly concentrated around 0.47. The departure of the null distribution of the posterior predictive p-values from a Uniform(0,1) distribution is nothing new, as discussed in Section 3.
Figure 4: Example of p-values being less variable than a \textit{Uniform}(0, 1) variable.

\textit{Case: Data Model 2PL or 3PL; Analysis Model 1PL}

The Rasch model is the simplest of the IRT models and makes restrictive assumptions (of equality of the item discriminations and no guessing)—it should be relatively easy to find powerful discrepancy measures under this situation.

\textit{Observed Score Distribution}

The observed score distribution seems to be somewhat useful in detecting the inadequacy of the Rasch model in these situations, especially when data are generated from the 3PL model. Consider one data set generated from the 3PL model and analyzed using the Rasch model. Figure 5 provides a graphical summary of the outcome of the posterior predictive checks for the data set with the observed score distribution as the discrepancy measure. For each possible score in the test (any integer between 0 and 30), a boxplot depicts the (empirical) posterior predictive distribution of the number of examinees obtaining that particular raw score. The boxplot shown is a little variation from the standard boxplot in the sense that the whiskers of the boxes extend to the 5th and 95th percentiles of the distribution. A notch towards the middle of each box denotes the median. The points indicate the observed number of people getting a particular score; a solid line joins the points for ease of viewing.

As an example, for score 7, the observed number of examinees is 30 while the median of the predicted (replicated) numbers is 24 and the 5th and 95th percentiles are 16
Figure 5: Boxplots for observed score distributions when the data model is 3PL and the analysis model is 1PL.

... to 32, respectively. The corresponding PPP-value is 0.10. The plot shows that towards the high end and low end of the score distribution, the 1PL model poorly predicts the number of people at a particular score. There are a number of cases in the middle as well where the Rasch model performs poorly, e.g., for score 14—the observed number is 117 while the 95th percentile of the replicated numbers is only 113 (PPP-value of 0.026)—and scores 11 and 16 (PPP-values 0.038 and 0.951).

A similar phenomenon occurs for the other 99 generated data sets as well. Figure 6 shows boxplots for the PPP-values for the 100 data sets together. The width of the boxplot for any specific observed score is proportional to the mean number of people (over the 100 data sets) for that score. There are horizontal dotted lines at 0.05, 0.5, and 0.95 to help interpret the results. The plot shows very interesting patterns. The PPP-values for the lower end of the observed score range are consistently high, which means that the Rasch model overpredicts the number of people with a low observed score. The Rasch model does
Figure 6: Boxplots for PPP-values corresponding to the observed score distributions when the data model is 3PL and the analysis model is 1PL.

not have a guessing parameter and hence assumes that an examinee with low proficiency will get all items wrong—that maybe the reason that it predicts more examinees for low score-points. It is difficult to explain the overprediction at the high end of the score range (we observed the same phenomenon for analysis of one data set). The fluctuating nature of the PPP-values in between is probably an outcome of the inherent difference between the ICCs of the Rasch model and the 3PL model. Probably the item discriminations play a role in these fluctuations as well.

Figure 7 shows similar boxplots for data generated from a 2PL model and analyzed using a Rasch model. The figure indicates a somewhat similar pattern as the previous figure.

Figure 8 shows similar boxplots for data generated from a Rasch model and analyzed using a Rasch model. The difference of Figure 8 (when the analysis model is the same as the data model) from Figures 6 and 7 (when the analysis model is not the data model) is very clear. However, the tendency to overpredict the numbers for the high scores even under the true model assumption partly explains the same type of pattern for data generated from 2PL and 3PL models.

It is obvious that the posterior predictive checks using the observed score
Figure 7: Boxplots for PPP-values corresponding to the observed score distributions when the data model is 2PL and the analysis model is 1PL.

Figure 8: Boxplots for PPP-values corresponding to the observed score distributions when the data model is 1PL and the analysis model is 1PL.

distribution as a discrepancy measure have some usefulness to detect the inadequacy of the Rasch model.
still underestimated the variability in the observed correlation coefficients. The situation that occurs when data come from a 2PL model is no different—the Rasch model replaced values closely indicates a limitation of the Rasch model for this data set. The observed standard deviation of the observed correlation coefficients is shown in Figure 10. This figure shows the difference between the standard deviation of the Rasch model and the observed correlation coefficient. It is straightforward to find a single PPF-valued interval that the observed correlation coefficient is close to the same discrimination parameter. It is also straightforward to find a single PPF-valued interval that the observed correlation coefficients are all very close together, around 0.6. This is not surprising as the Rasch model assumes all items to have the same discrimination parameter.

More importantly, the replaced values are all very close together, around 0.6. This is not surprising as the Rasch model assumes all items to have the same discrimination parameter. It is also straightforward to find a single PPF-valued interval that the observed correlation coefficients are all very close together, around 0.6. This is not surprising as the Rasch model assumes all items to have the same discrimination parameter.

Figure 9: Boxplots for observed correlation coefficients when the data model is 3PL and the analysis model is 1PL.

The figure shows a boxplot of the replaced observed correlation coefficients for each item. The box extends from the 25th percentile to the 75th percentile, which are the first and third quartiles. The line inside the box indicates the median. The whiskers extend from the box to the 10th and 90th percentiles. The dots denote outliers, which are outside the range of the whiskers. The data show a clear trend of the observed correlations being closer to the 3PL model than the 1PL model.
**Figure 10:** Histogram for standard deviation of biserial correlations when the data model is 3PL and the analysis model is 1PL.

The same phenomenon of underestimation (of the variability in biserial correlations) by the Rasch model occurs in the other 99 data sets as well and we do not show the results.

**Interaction/Association Among the Items**

The odds ratio measure seems to be quite powerful in detecting the inadequacy of the Rasch model. The additive version of the odds ratio has considerable power as well; but as described earlier, the measure provides the same information as the odds ratio and has comparatively less power—so we will discuss results for the odds ratio only.

Consider the case when the generating model is the 3PL model. The top left panel of Figure 11 shows boxplots of the replicated odds ratios for all item-pairs involving item 1, i.e., the pairs (1,2), (1,3), (1,4), ..., (1,30). The other panels in the figure show similar plots for items 2, 15, and 22. The dots show the observed odds ratios for each pair. Consider the plots for Item 1—there is one box for each item pair possible involving Item 1. The labels of the horizontal axis indicate the other item in the pair. For example, the replicated odds ratios for the item pair (1, 2) are all around 2 while the observed value for the pair is slightly less than 2.

The plot reveals a number of interesting facts. First, the replicated odds ratios all center around a little above 2. This equality of all predicted odds ratios follows intuitively from arguments in Holland (1990a), although a complete proof is yet to be found (personal communication with Paul Holland). Intuitively, the Rasch model assumes the same
Figure 11: Boxplots for observed and replicated odds ratios when the data model is 3PL and the analysis model is 1PL.

discrimination parameter for all the items and hence has all item response functions parallel to each other, causing the predicted odds ratios to be very close to each other.

Next, by comparing the observed odds ratios to the replicated odds ratios, it is obvious that the Rasch model does an inadequate job of predicting the associations among the items. The observed values are almost always greater than the replicated values for Items 2 and 15 (items with medium/high discrimination), while the opposite is true for
Items 1 and 22 (item with low discrimination). The plot for the other items show similar patterns.

The above results are for one data set in the simulation study. The other 99 data sets (when data are from 2PL/3PL model) show similar results. Figure 12 summarizes the median (from all 100 data sets) PPP-value for the odds ratio measure for each pair of items when data are generated from a 2PL model. Items are sorted (in an increasing order) according to their true discrimination parameters, so that the symbol for (1, 2) corresponds to the pair with the lowest discriminating item and the second-lowest discriminating item. A solid triangle for an item pair indicates that the median of the 100 PPP-values for that

![Plot](attachment:image.png)

**Figure 12**: Median PPP-values for odds ratios when data model is 2PL and analysis model is 1PL.

pair is above 0.95 while a hollow inverted triangle for a pair indicates that the median is below 0.05. No symbol for an item pair means that the median PPP-value is within 0.05 and 0.95. The legend for the figure describes the symbols. Because the odds ratio (and hence the PPP-value) for the item pair (1, 2) is the same as that for the pair (2, 1), the plot
shows only one point for each item pair, resulting in a triangular array of points. Horizontal and vertical grid-lines at multiples of 5, vertical axis-labels on the right side of the figure, and a line joining the points (0,0) and (30,30) are there for convenience.

The plot shows that the Rasch model cannot adequately explain the associations among the items. The plot also shows a clear pattern; the model overestimates the associations involving the low-discriminating items, while underestimating the associations involving the most difficult items. The Rasch model assumes that the level of association is the same for all item pairs; however, the data, coming from a 2PL model, has lower association for pairs involving one low-discrimination item (which do not contribute much information) and higher association for other pairs.

Figure 13 provides an idea about the power (effectiveness in detecting the model inadequacy) of the PPP-values. In practice, a researcher has only one data set to analyze and often the judgment on the model fit is based on an extreme p-value—hence it is important to examine how often one will observe an extreme p-value or a data set. Figure 13 provides such information. The plot summarizes the proportion of times (out of 100) the PPP-values are extreme (below 0.05 or above 0.95). If the proportion lies between 0.95 and 1, the color of the symbol for that pair in Figure 13 will be that of the leftmost strip in Figure 14. If the proportion is between 0.9 and 0.95, the color will be that of the next strip. In essence, the darker a point for a pair, the higher is the proportion of times the PPP-value is extreme for that pair.

Figures 12 and 13 tell the same story—whenever there is a triangle in the former (indicating an extreme median), the corresponding point in the latter figure is black.

Hence, the odds ratio performs quite well in indicating the inadequacy of the Rasch model to predict the interaction between the items. When the generating model (data model) is the 3PL model, the situation is no different—the odds ratios are still very powerful (and more powerful than its additive version) in indicating the inadequacy of the Rasch model. Figure 15 shows a plot similar to Figure 12 for this case. The model still overestimates the odds ratio for item pairs with one low-discriminating item and underestimates the odds ratios for other pairs. The only exception are the pairs involving Item that has rank 22 with respect to discrimination; that specific item also happens to be
Figure 13: Power of PPP-values for odds ratios when the data model is 2PL and the analysis model is 1PL.

a very difficult item.

Incidentally, the median PPP-value plot for the odds ratios when data is generated from a Rasch model and analyzed using the Rasch model has no triangle at all. Hence, the odds ratio measure has low type I error and high power in assessing the fit of the Rasch model.

The above results show that it is easily possible to detect the inadequacy of the Rasch model when the generating model is either a 2PL or 3PL model. Therefore, this report does not show any results for fitting the Rasch model to more complicated models.

Figure 14: The color range used.
Figure 15: Median PPP-values for odds ratios when the data model is 3PL and the analysis model is 1PL.

(e.g., the testlet model)—the same measures found powerful here work in those situations as well.

Case: Data Model 3PL; Analysis Model 2PL

The 2PL model is an extension of the Rasch model with the added assumption of different discrimination parameters for the items. However, the model does not have a guessing parameter as does a 3PL model. Hence, it will be of interest to examine how the discrepancy measures perform when the 2PL model is fitted to data generated from the 3PL model. A number of experts believe that the 2PL model usually fits psychometrics data significantly better than the 1PL model, but not significantly worse than the 3PL model (see, e.g., Orlando & Thissen, 2000; Yen, 1981). Therefore, keeping in mind the conservativeness of the PPCMC method, it may not be straightforward to find powerful discrepancy measures under this situation.
**Observed Score Distribution**

As before, let us examine the results when one data set simulated from the 3PL model is analyzed using the 2PL model. Figure 16 shows boxplots for the distribution of

![Observed Score Distribution](image)

*Figure 16: Boxplots for observed score distributions when the data model is 3PL and the analysis model is 2PL.*

the replicated number of examinees (out of 2500) and points for the observed numbers for each score, a solid line joining the points. The plot shows some patterns similar to Figure 5, e.g., the 2PL model over-predicts the number of people at the low end of the score distribution and does not seem to predict satisfactorily the numbers towards the middle either—e.g., the observed numbers are larger than the 3rd quartile for all scores but one within 10 to 15. However, the figure does not show any severe problems with the model.

A similar phenomenon occurs for the other 99 generated data sets as well. Figure 17 shows boxplots for the PPP-values for the 100 data sets together. The width of the boxplot for any specific observed score is proportional to the mean number of people (over the 100 data sets) for that score. There are horizontal dotted lines at 0.05, 0.5, and 0.95
to help interpret the results. The plots show patterns very similar to those in Figure 6

![Boxplots for PPP-values for observed score distributions when the data model is 3PL and the analysis model is 2PL.](image)

**Figure 17**: Boxplots for PPP-values for observed score distributions when the data model is 3PL and the analysis model is 2PL.

except that for the high scores, the 2PL model seems to underpredict the number of people getting those scores. Also, the PPP-values seem to be more extreme than those in Figure 6. The effect of the lack of a guessing parameter is strong enough for the observed score discrepancy measure to detect the apparent inadequacy of the 2PL model.

**Biserial Correlation Coefficients**

Figure 18 shows a boxplot of the replicated biserial correlation coefficients for each item for one data set. The dots denote the observed biserial correlation coefficient for each item. The 2PL model does much better than the Rasch model in predicting the biserial correlations—the dot lies within the 90% credible intervals for all the items. However, careful examination of the plot suggests that for a number of items, the 2PL model seems to slightly overpredict the biserial correlations.

Figure 19 shows histograms of the replicated values of the mean and standard deviation of the biserial correlation coefficients. Vertical lines show the observed values. The histograms show that the 2PL model overpredicts both the mean and variance of the
**Figure 18**: Boxplots for biserial correlations when the data model is 3PL and the analysis model is 2PL.

**Figure 19**: Histogram for mean and standard deviation of biserial correlations when the data model is 3PL and the analysis model is 2PL.

Biserial correlations. The effect with the variance is more severe (PPP-value 0.95). Further examination reveals that the 2PL model underpredicts the minimum biserial correlation (PPP-value 0.25) and overpredicts the maximum biserial correlation (PPP-value 0.92), with overpredicting the variance as an end-result. However, these are not overwhelming evidence of inadequacies of the 2PL model.

The situation is no different in the other 99 data sets as well. Figure 20 shows histograms of the PPP-values for the mean and standard deviation of the biserial
correlations. The vertical lines show the median PPP-value.

![Graph showing Average and SD](image)

**Figure 20:** PPP-values for mean and standard deviation of biserial correlations when the data model is 3PL and the analysis model is 2PL.

Probably the lack of a guessing parameter makes the predicted item discrimination powers higher and more variable in the 2PL model than in the 3PL model, resulting in the overprediction of the mean and variance of the biserial correlations by the 2PL model.

**Interaction/Association Among the Items**

The 2PL model reproduces adequately the odds ratios (and the additive version of them) under this situation. A plot of the median PPP-values for the odds ratio discrepancy measure from the 100 data sets combined has no median above 0.95 or less than 0.05 at all. Figure 21 summarizes the power of the PPP-values for the odds ratio discrepancy measure from the 100 data sets, using the same symbols as in Figure 13.

The results shows that the odds ratios have very little power to indicate any inadequacy of the 2PL model when the data generating model is the 3PL model.

Hence, the discrepancies considered in this work do not suggest any serious misfits when the data model is the 3PL model and the analysis model is the 2PL model. That is probably due to a combination of the facts that the posterior predictive checks are conservative in nature and that the 2PL model explains adequately data from a 3PL model—at least the aspects of the data addressed by the discrepancies used. One limitation here is that the simulation study does not consider a very large data set—the same discrepancy measures may find the 2PL model performing inadequately for such a data set.
Figure 21: Power of PPP-values for odds ratios when the data model is 3PL and the analysis model is 2PL.

We will observe later that a 2PL model performs relatively worse than a 3PL model in one of our real data examples, where the data set is large.

Case: Data Model Multidimensional; Analysis Model 2PL/3PL

This part aims at finding discrepancy measure(s) having enough power to detect any lack of unidimensionality in the data. For example, if different items in a test measure different skills, the test is a multidimensional test and no unidimensional model should be able to explain the data adequately. It is of interest here to examine if a 2PL or a 3PL model can explain adequately data in such a situation.

For generating data from the linear logistic two-dimensional model (5), the data generator uses the difficulty and guessing parameters shown in Table 2. The ability parameters for the two dimensions, $\theta_{1j}$ and $\theta_{2j}$, are generated from bivariate normal $N_2(0, 0, 1, 1, \rho)$ distributions for different values of $\rho$. 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 $a_j$ in Table 2 as:

\[ a_{1j} = a_j I_{[j \leq 15]}, \quad a_{2j} = a_j I_{[j > 15]}, \]

38
Therefore, by the data generation scheme, the first 15 items in the test measure one skill while the last 15 items measure another skill.

**Extreme Multidimensionality**

First, we generate 100 data sets under the assumption that the abilities of the two dimensions are independent of each other, that is, the correlation $\rho$ between $\theta_{1i}$ and $\theta_{2i}$ is 0, i.e., under extreme multidimensionality. The 3PL model severely underpredicts the biserial correlations for the first 15 items while severely overpredicting the biserial correlations for the last 15 items (output not shown) for all the 100 data sets.

Figure 22 shows boxplots for the PPP-values corresponding to the observed score distribution for the 100 data sets. The plots show clear patterns of underprediction (in the ends of the score range) and overprediction to indicate serious misfit of the 3PL model.

![Boxplots for PPP-values](image)

**Figure 22:** Boxplots for PPP-values for observed score distributions when the data model is two-dimensional (with uncorrelated dimensions) and the analysis model is 3PL.

Figures 23 and 24 summarize the median and power of the PPP-values for the odds ratio discrepancy measures (using the same symbols used for Figures 15 and 13) from the 100 data sets. The figures show that the PPMC method clearly points to some association among the items in the test. The 3PL model severely underpredicts all the odds ratios
**Figure 23:** Median PPP-values for odds ratios when the data model is two-dimensional (with uncorrelated dimensions) and the analysis model is 3PL.

**Figure 24:** Power of PPP-values for odds ratios when the data model is two-dimensional (with uncorrelated dimensions) and the analysis model is 3PL.

involving the first 15 items only, revealing the two-dimensional nature of the items.
**Moderate Multidimensionality**

This paper also examines the case with less extreme (and more realistic) multidimensionality by varying the correlation coefficient between $\theta_{1i}$ and $\theta_{2i}$ and performing posterior predictive checks.

Let us consider the case when the correlation coefficient between $\theta_{1i}$ and $\theta_{2i}$ is 0.6. The 3PL model predicts the biserial correlations and observed score distribution more or less satisfactorily. The story is different with the odds ratios, however. Figure 25 summarizes the medians of the PPP-values for the odds ratios from 100 data sets in this situation. The figure shows that the odds ratios have enough power even in this situation. However, the figure differs from Figure 23 in that it has extreme medians for both sets of items, 1-15 and 16-30, making it little more difficult to detect that the test is two-dimensional. However, careful examination reveals that the 3PL model mostly underpredicts the odds ratios within the two groups. The effect is much more severe for the
first set, with a significant number of black inverted triangles. These underpredictions are outcomes of associations among the items beyond the model’s prediction capacity.

On the other hand, the model overpredicts the association on many instances for the pairs involving one item from each set. The 3PL model assumes that all examinees at a specific model-predicted proficiency level should get all items correct within the reach of their level. However, some of the examinees at the same model-predicted proficiency level actually have higher proficiency for one dimension and lower proficiency for the other dimension in the data model. For an item pair with one item each from the sets 1-15 and 16-30, some of the above-mentioned examinees will be likely to get one item correct and the other wrong—the end-result is an observed odds ratio lower than the predicted (and hence a high p-value) for that item pair. The plot showing the power of the PPP-values (not shown) provides the same information.

**Case: Data Model Testlet Model; Analysis Model 3PL**

For generating data from the testlet model (4), we assume that there are six testlets with five items in each testlet. Items 1 through 5 belong to Testlet 1, Items 6 to 10 belong to Testlet 2, and so on. The generating item parameters are the same as those on Table 2. The program tried several sets of testlet variances $\sigma_{d(j)}^2$, as discussed later. Testlet effects $\gamma_{id(j)}$s are generated from their distribution $\gamma_{id(j)} \sim N \left(0, \sigma_{d(j)}^2\right)$.

Consider the case when the data-generating model is the testlet model (4) with six testlets and corresponding testlet standard deviations $\sqrt{\sigma_{d(j)}^2}$ of 0, 0.3, 0.5, 0.7, 0.85, and 1.

The observed score distribution and the biserial correlations fail to detect any inadequacy of the 3PL model—plots for these measures look very similar to those when data are from the 3PL model itself. Fortunately, the odds ratios are powerful in this situation. Figures 26 summarizes the median the PPP-values for the odds ratio discrepancy measures from the 100 data sets combined.

The figure (along with the plot, not shown here, for the power of the PPP-values) shows that the PPMC method can detect the testlet structure of the test. The 3PL model underpredicts the odds ratios, which are the measures of associations, among the items in the last three testlets (Items 16-20, 21-25, and 26-30). The model underperforms for part
Figure 26: Median PPP-values for odds ratios when the data model is the testlet model and the analysis model is 3PL.

of the third testlet, consisting of Items 11-15, as well. Interestingly, the model appears to explain the odds ratios for the first two testlets (Items 1-5 and 6-10) adequately. This is not very surprising, considering that the generating testlet variances for these testlets are 0 and 0.09, making the items in those effectively ones from a 3PL model.

This work also considers other values of testlet variances to generate testlet data. Those analyses show that the 3PL model can predict the odds ratios adequately when the maximum testlet variance is low. For the settings here (2,500 examinees, 30 items and item parameters as shown in Table 2), the model seems to predict the odds ratios satisfactorily if all the generating testlet variances are less than 0.25. About one particular testlet, whether the model can explain the odds ratios involving items within the testlet depends on the variance for that testlet and the variance for the other testlets as well—but the model mostly succeeds if the testlet variance is less than about 0.3 in our limited simulation studies.
**Case: Data from Speededness Model; Analysis Model 3PL**

This document will use the term *speededness model* to denote a data-generating model that generates data assuming that the test is speeded. To generate data from such a model, we take the item response function to be the 3PL model, but assume that a certain percentage of the examinees are affected by the time limit towards the end of the test. The percentage of examinees affected vary in different simulation studies. A random integer drawn uniformly from the set \{23, 24, \ldots, 30\} denotes the item from which an examinee gets rushed, if at all, and scores 0 in all the remaining items (i.e., gets the item wrong or omits it). This data generation scheme maybe somewhat unrealistic (considering that a number of experts think that speededness occurs gradually, i.e., the performance of the examinees deteriorate as they get rushed—examinees do not start answering incorrectly from a certain time-point), but is used for its simplicity.

**Extreme Speededness**

First, let us simulate the extreme case when all the examinees are affected by the time limit. If the PPMC method cannot detect misfit under this extreme speededness, there is no hope of success with the method under realistic, less extreme situations.

*Observed score distribution.* Figure 27 shows boxplots for the distribution of the replicated number of examinees (out of 2500) and points for the observed numbers for each score for one data set. A solid line joins the observed scores. The plot shows clear signs of misfit of the 3PL model in this situation. The model overpredicts the number of people at the high end of the score distribution, and does not seem either to predict satisfactorily the numbers from the middle until the right end.

A similar phenomenon occurs for the other 99 generated data sets as well. Figure 28 shows boxplots for the PPP-values for the 100 data sets together. The width of the boxplot for any specific observed score is proportional to the mean number of people (over the 100 data sets) for that score. There are horizontal dotted lines at 0.05, 0.5, and 0.95 to help interpret the results. The plots show clear fluctuating patterns of underprediction and overprediction to indicate serious misfit of the 3PL model.
Figure 27: Boxplots for observed score distributions when data model is speededness (extreme) model and analysis model is 3PL.

Figure 28: Boxplots for PPP-values for observed score distributions when the data model is the speededness (extreme) model and the analysis model is 3PL.
Biserial correlations. The model severely overpredicts the biserial correlations for the last four items and moderately underpredicts the same quantities for a number of other items (even those that are not speeded), indicating some problems with the model.

Interaction/association among the items. Figure 29 summarizes the median of the PPP-values for the odds ratios from the 100 data sets combined. The figure shows that the

Figure 29: Median PPP-values for odds ratios when the data model is the extreme speededness model and the analysis model is 3PL.

PPMC method clearly points to some problems with the model. The plot of the power of the p-values (not shown) tell the same story. The 3PL model mostly underpredicts the odds ratios for pairs involving speeded items only. This is expected, keeping in mind the way of generation of the data sets—the data generator assumes that when rushed, an examinee will get the last few items wrong, thus introducing something similar to a testlet effect for the last few (speeded) items in the test. On the contrary, the model mostly overpredicts the odds ratios involving a speeded item and a non-speeded item. The 3PL model assumes that
all examinees at a specific proficiency level should get all items correct within the reach of their level. However, for an item pair with a speeded item and a nonspeeded item, some of the examinees rush on the speeded item and get that incorrect—the end-result is an observed odds ratio lower than the predicted (and hence a high p-value) for that item pair.

**Moderate Speededness**

Let us simulate the case when 20% examinees are affected by the time limit and get all the items from a randomly generated number between 23 and 30 wrong, which is a more realistic situation. For example, Schnipke and Scrams (1997) find that in the 1992-93 GRE general analytical section, which has 25 questions, the percentage of examinees not reaching Items 20 to 25 grow step-by-step from 2.5 to 15. Also, 17% to 50% of examinees engage in rapid-guessing behavior in the last 10 items; The percentages of examinees not reaching Items 23 to 30 in this simulation study grow from 2.5 to 20, which is quite at par with the findings of Schnipke and Scrams (1997).

The biserial correlations do not show any problems with the model. The predicted observed score distributions show (plot not provided) some inadequacy of the 3PL model, but the plots look much less extreme than Figure 28.

Figures 30 summarizes the median PPP-values for the odds ratio discrepancy measure from the 100 data sets combined. It is clear from the figure that the odds ratios have some power to indicate problems with the model even in these circumstances, although the power is less than that under extreme speededness.

**Case: Data from Multidimensional and Speededness Model; Analysis Model 3PL**

We generate data from the two-dimensional model (5) with $\rho=0.6$, i.e., under moderate multidimensionality, and assume that there also exists moderate speededness (as described earlier), 20% examinees being affected on Items 23-30. Figures 31 summarizes the median PPP-values for the odds ratio. The figure shows that the odds ratios seem to successfully detect the nature of violations of the assumptions in the 3PL model. In a way, the above figure has some similarity to Figures 25 and 30.
Figure 30: Median PPP-values for odds ratios when the data model is the speededness (moderate) model and the analysis model is 3PL.

A Brief Summary of the Simulation Studies

The simulation studies prove that the posterior predictive model checking method has the potential to detect the misfit of common IRT models for a variety of circumstances.

Table 3 summarizes the findings, from our simulation studies, regarding the performances (power) of the discrepancy measures. The table describes the power of the measures, in a scale of 0-2 (where 0 means no/little power, 1 means some power, and 2 means considerable power), to detect different types of model misfits. Because the work finds that the percentage-correct score for items has no power in any of the situations considered, the measure does not appear in the table.

The table indicates that the odds ratios are very powerful in assessing different aspects of misfit of the IRT models, especially those related to lack of unidimensionality. The observed score distribution and the biserial correlation coefficients are useful too on some occasions.
Figure 31: Median PPP-values for odds ratios when data come from a moderately multidimensional and moderately speeded test, but the data are analyzed using a 3PL model.

7. Conclusions

In a real application, a researcher should always assess the fit of the statistical model used. Researchers van der Linden and Hambleton (1997) observe the lack of well-established model-checking tools for models in item response theory.

As this work shows, posterior predictive model checking provides a straightforward way to perform a collection of model checks, aimed at different aspects of the model. The simulation studies show that the PPMC methods have a low type I error (i.e., chance of detecting a problem wrongly) and considerable power (chance of detecting a problem rightfully) in detecting the misfit of an IRT model. A companion article (Sinha et al. & Johnson, in press), discussing applications of the method to real data examples, further supports the findings from these simulations. One strong aspect of this report is the use of easily comprehensible and attractive graphical displays to demonstrate the strength of the
Table 3: *The Power of the Discrepancy Measures to Detect Different Types of Model Misfits (0: no power, 1: some power, 2: considerable power)*

<table>
<thead>
<tr>
<th>Type of misfit addressed</th>
<th>Obs. score disn</th>
<th>Biserial corr.</th>
<th>Percent correct for item pairs</th>
<th>Odds ratio (&amp; its additive version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inadequacy of a Rasch model for data from 2PL/3PL model</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Inadequacy of a 2PL model for data from 3PL model</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Inadequacy of a 3PL model for 2-dimensional data</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Inadequacy of a 3PL model for data from testlet model</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Inadequacy of a 3PL model for speededness data</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

PPMC method. It is clear from this work that the PPMC method can successfully detect different aspects of misfit of the common IRT models in a number of different situations. Thus, this report shows that the technique needs more attention of the psychometricians.

Another positive aspect of this work is that it provides a better understanding of the IRT models, especially of their ability to predict a number of observed quantities (like the observed score distribution, the odds ratios, etc.) of natural interest related to test data. For example, this work finds that the Rasch model performs quite poorly in reproducing the interaction effects among the items.

The choice of discrepancy measures is a vital issue in the application of the PPMC methods. A number of discrepancy measures examined in this work appear promising in assessing IRT model fit. The odds ratios, examining the association between item pairs, are found to be a very powerful discrepancy measure. The measure seems to be especially powerful in detecting the lack of unidimensionality in the data. The measure can also detect any significant speededness in the data. In the simulation studies, the odds ratio measure detects the different types of violations of unidimensionality in the data sets. The
real applications (in a companion article) provide further support of the power of the odds ratios as discrepancy measures.

The PPMC method is time-consuming, especially compared to a classical analysis. However, a fact that gives the method an edge is that standard classical tests have not found many applications in psychometrics yet (even though they may, in principle, be applied). Once a model is fit using an MCMC algorithm and the posterior sample is stored, the PPMC method can check virtually any aspect of the fit of the model with no additional theoretical work and a little extra computational effort (whereas, to do the same in a classical approach, a researcher will have to face the often hard and occasionally impossible task of establishing the null distribution of a test statistic). Further, application of the classical methods will be very difficult, even impossible to the recently developed rather complicated models (e.g., Bradlow et al., 1999; van Onna, 2003; Fox & Glas, 2003), making the PPMC method the only possible model checking tool for them. No wonder there has been a recent surge in the use of the PPMC method in psychometrics, as a list of applications of the technique in Section 1 shows.

However, there are a number of unresolved issues that require attention in the future. Some of them are: (a) examining item fit measures, (b) examining realized discrepancies (Gelman et al., 1996), those involving data as well as parameters, (c) extending the work to take into account polytomous items, and (d) exploring the possibilities of person fit analyses, etc.
References


Pearson, K. (1909). On a new method of determining a correlation between a measured character of A and a character of B, of which only the percentage of cases wherein B exceeds (or falls short of) intensity is recorded for each grade of A. *Biometrika, 7*, 96-105.
Copenhagen, Denmark: Danish Institute for Educational Research.


