Bayesian Item Fit Analysis for Dichotomous Item Response Theory Models

Sandip Sinharay
Bayesian Item Fit Analysis for Dichotomous Item Response Theory Models

Sandip Sinharay

ETS, Princeton, NJ

December 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
ETS
Princeton, NJ 08541
Abstract

Detecting item fit for common dichotomous item response theory (IRT) models has always been an issue of enormous interest, but there exists no unanimously agreed upon item fit diagnostic for the models. This paper employs the posterior predictive model checking method (Guttman, 1967; Rubin, 1981, 1984), a popular Bayesian model checking tool, to examine item fit for common dichotomous item response theory models. An item fit plot, comparing the observed and predicted proportion correct scores of examinee groups (with groups based on raw scores), promises to be useful in real applications. This paper also suggests how to obtain posterior predictive p-values for the $\chi^2$-type test statistics of Orlando and Thissen (2000) comparing the observed and predicted proportion correct scores for different raw score groups. A number of simulation studies and real data applications examine the effectiveness of the suggested item fit diagnostics. The suggested p-values seem to have low Type I error rate, low false alarm rate, and adequate power, indicating that researchers might find the methods more acceptable than the existing item fit measures.

Key words: Discrepancy measure, item response theory (IRT), posterior predictive model checking, p-values
1. Introduction

Detecting item fit has always been an issue of enormous interest in psychometrics. For the common dichotomous item response theory (IRT) models, one way to detect item fit is to compare the average item performance (e.g., number of correct answers/proportion correct score) of various groups of examinees to the performance predicted by the model. Hambleton and Swaminathan (1985) and Hambleton (1989) suggest using graphical plots to make the comparison.

Bock (1972), Yen (1981), McKinley and Mills (1985), Stone, Mislevy, and Mazzeo (1994), and Donoghue and Hombo (1999, 2003) suggest test statistics that quantify the comparison. They use groups based on the estimates of ability parameters of the examinees. However, grouping on the basis of model-estimated ability (that is not really observed) of the examinees is not in the same spirit as in a traditional $\chi^2$ goodness-of-fit test. Another problem with these test statistics is that the $\chi^2$ approximation of them does not hold (because of estimating the “observed” quantities) and hence, their null distributions are not clearly established.

Orlando and Thissen (2000, 2003) suggest two $\chi^2$-like test statistics (known as $S - \chi^2$ and $S - G^2$) after grouping the examinees based on their number-correct (raw) scores. This idea of grouping is similar to that used for Rasch model diagnostics (see, e.g., Molenaar, 1983). This approach is more intuitive and acceptable than those using groups based on estimated abilities because the approach deals with truly observed counts. Glas and Falcon (2003) suggest an item fit test that is based on the Lagrange multiplier test (or equivalent efficient score test) and uses examinee groups based on number-correct scores. However, Glas and Falcon (2003) find the overall characteristics of their suggested item fit test to be worse than the characteristics of the tests suggested by Orlando and Thissen (2000), making the latter the most acceptable item fit tests as of now.

However, using arguments from Chernoff and Lehmann (1953), it is possible to prove that the test statistics suggested by Orlando and Thissen (2000) do not follow limiting $\chi^2$-distributions but are stochastically larger than the assumed $\chi^2$ random variables. The departure maybe severe for short tests. In summary, none of the existing approaches for
assessing item fit seem to be adequate and further research needs to be done 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 obvious intuitive appeal. The method primarily consists of 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 discrepancy measures for the observed data set and those for 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.

This paper uses the PPMC method to examine item fit for simple dichotomous IRT models. As in Orlando and Thissen (2000), the raw scores of the examinees are the basis of forming the examinee groups. An item fit plot, showing the observed and predicted proportion correct scores for the examinees for different groups, provides information on the groups for which the observed and predicted proportions differ significantly. We also use two discrepancy measures based on the $\chi^2$-type measure and the $G^2$-type measure suggested by Orlando and Thissen (2000). The posterior predictive p-value (PPP-value) corresponding to the discrepancy measure quantifies the information from the item fit plots.

Section 2 introduces the common IRT models and then discusses the existing item fit measures. Section 3 provides a brief discussion of the PPMC method and the posterior predictive p-value. Section 4 introduces the discrepancy measure and the item fit plot and then discusses how to compute the item fit p-values for the discrepancy measure. A number of simulation studies examine the effectiveness of the item fit plots and item fit p-values in Section 5. A second set of simulation studies in Section 6 demonstrates that the PPP-values appear to have satisfactory power, but low Type I error rates and false alarm rates; they do not flag items unjustifiably (unlike the measures suggested by Orlando and Thissen, 2000; Glas and Falcon, 2003). Section 7 applies the suggested measures to a
number of real data examples; many will find these examples useful due to the lack of many applications of item fit measures to real data sets (e.g., there is no published application of the measures suggested by Orlando and Thissen, 2000; Glas and Falcon, 2003, to real data sets yet). Section 8 discusses the conclusions of this study.

2. Dichotomous Item Response Theory Models and the Existing Item Fit Tests

The Models

Consider an educational assessment, consisting of \( I \) items, that is given to \( J \) individuals. Let \( y_{ij} \) denote the score of the \( j \)-th individual to the \( i \)-th item; the score is 1 if the examinee answered the item correctly and 0 otherwise.

The three-parameter logistic (3PL) model (Lord, 1980) assumes that

\[
P(y_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i) \logit^{-1}(a_i(\theta_j - b_i)), \tag{1}
\]

where \( \logit^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} \), \( a_i \), \( b_i \), and \( c_i \) respectively are the slope, difficulty and guessing parameters of item \( i \), and \( \theta_j \) is the proficiency/ability parameter for examinee \( j \).

The Rasch model (Rasch, 1960), also referred to as the one-parameter logistic (1PL) model, is given by

\[
P(y_{ij} = 1 \mid \theta_j, a_i, b_i) = \logit^{-1}(a_i(\theta_j - b_i)), \tag{2}
\]

and the two-parameter logistic (2PL) model (Birnbaum, 1968), is given by

\[
P(y_{ij} \mid \theta_j, a_i, b_i) = \logit^{-1}(a_i(\theta_j - b_i)) \cdot \tag{3}
\]

The above three models are the most popular models in psychometrics and the testing companies around the world use these models operationally.

In this paper, a standard normal population distribution assumption,

\[
\theta_j \overset{iid}{\sim} \mathcal{N}(0, 1), \quad j = 1, 2, \ldots J
\]

fixes the scale for the 1-, 2-, and 3PL model. Denote \( \eta_i \) to be the vector of item parameters for item \( i \). This paper fits the model (as in Holland, 1990).

\[
\int_{\theta} \prod_{i=1}^{I} P(y_{ij} \mid \theta, \eta_i)^{y_{ij}} (1 - P(y_{ij} \mid \theta, \eta_i))^{1-y_{ij}} \phi(\theta) d\theta
\]
to the response vector of examinee \( j, j = 1, \ldots, J \), treating the examinee proficiencies as
nuisance parameters, where \( \phi(\theta) \) denotes the ordinate of the \( \mathcal{N}(0,1) \) distribution at \( \theta \).

**A General Description of the Item Fit Tests**

The steps in computing any of the existing item fit measures for IRT models is as
follows:

1. Fit an IRT model to a data set to obtain estimates of ability parameters and item
parameters.

2. Form a number of examinee groups defined on the scale of \( \theta \) or on the scale of raw
scores.

3. Place each examinee into one of the groups according to his/her \( \theta \)-estimate or raw
score; the approach by Stone et al. (1994) and Donoghue and Hombo (1999) does not
place an examinee into only one group, but computes probabilities for an examinee to
belong to each group.

4. For each item and each examinee group, compute the *observed* number (proportion) of
examinees in the group who answered the item correctly/incorrectly.

5. Compute the *expected* proportions corresponding to the above observed proportions;
this step may require considerable amount of computations.

6. Compute the value of the test statistic(s) comparing the observed and expected pro-
portions.

Suppose the examinees are divided into groups \( k, k = 1, 2, \ldots K \). Denote the number
of examinees in group \( k \) as \( N_k, k = 1, 2, \ldots K \). Suppose the interest is in testing the fit of
item \( i \). Denote the observed and expected proportion of examinees in group \( k \) answering
item \( i \) correctly as \( p_{ik} \) and \( E_{ik} \) respectively. The existing item fit statistics (for item \( i \)),
except the one by Glas and Falcon (2003) are of the form

\[
\chi_i^2 = \sum_{k=1}^{K} N_k \frac{(p_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}
\]  \( (4) \)
or

\[ G_i^2 = 2 \sum_{k=1}^{K} N_k \left[ \ln p_{ik} \ln \left( \frac{p_{ik}}{E_{ik}} \right) + (1 - p_{ik}) \ln \left( \frac{1 - p_{ik}}{1 - E_{ik}} \right) \right]. \tag{5} \]

The test statistic suggested by Glas and Falcon (2003), based on the Lagrange multiplier test principle, is of the form

\[ LM_i = h_i' \sum_i h_i, \tag{6} \]

where each component of the vector \( h_i \) is the difference between the observed proportion correct and its posterior expectation for a score-group, computed at the maximum likelihood estimate of the parameters, and \( \Sigma_i \) is the estimated variance matrix of \( h_i \).

The item fit approaches then use a \( \chi^2 \) approximation of the test statistics in (4), (5), and (6), or perform extensive simulation studies in an attempt to obtain the asymptotic distribution of the test statistics.

**Tests Using Examinee Groups Based on Estimates of Proficiency Parameter**

Bock (1972), Yen (1981), and McKinley and Mills (1985) suggest forming the examinee groups based on joint maximum likelihood estimates of proficiencies and item parameters. The former two suggest a \( \chi^2 \)-type test statistic while the latter suggests a likelihood ratio \( G^2 \)-type statistic. In contrast, Stone et al. (1994) and Donoghue and Hombo (1999) use marginal maximum likelihood estimates (MMLE) of the item parameters and examinee groups based on the posterior distribution of \( \theta \). In this approach, an examinee does not belong to only one group, but belongs to all the groups with probabilities (weights) determined by the posterior distribution of the corresponding \( \theta \). This approach is mainly designed for nonlinear tests (where examinees may receive different tests, maybe of varying lengths or difficulty levels).

However, grouping on the basis of the estimated ability of the examinees (which are not observed) is not in the spirit of the traditional \( \chi^2 \) (or \( G^2 \)) goodness-of-fit test—the “observed” counts (proportions) for a \( \chi^2 \) test should be available before any model is fitted, which is not true for the above measures. Another major problem with the measures is that their distributions under the null hypothesis/model (i.e., when the model fits the data) are
not clearly established. The asymptotic $\chi^2$ approximation is obviously not correct because the quantities treated as “observed” by these methods are actually estimated, and because of the collapsing of the original responses, using arguments in Chernoff and Lehmann (1953). A recent theoretical result in Donoghue and Hombo (2003) holds promise, but its application is computation-intensive and the result is not usefully implemented in real data sets yet. The third disadvantage with these methods is that there is no natural way to determine the number of groups of examinees. For example, Yen (1981) and McKinley and Mills (1985) use 10 groups, Donoghue and Hombo (1999) use 42 groups, and Bock (1972) uses a variable number of groups.

**Tests Using Examinee Groups Based on the Raw Scores**

For Rasch models, the number-correct (raw) score of an examinee is a sufficient statistic for the examinee ability; as an outcome, a number of Rasch model diagnostics (see, e.g., Molenaar, 1983) use groups of examinees based on their raw scores. Orlando and Thissen (2000, 2003) use the idea to suggest item fit tests for common dichotomous IRT models. In an $I$-item test, total number of possible groups is $(I + 1)$, one each for raw score $0, 1, 2, \ldots I$. The computation of the expected proportion of examinees for group (raw score) $k$ who answer item $i$ correctly uses the MMLEs of the item parameters and a recursive approach suggested by Lord and Wingersky (1984). The next step is to apply the $\chi^2$ approximation to (4) and (5). The summation in the test statistics ranges from $k = 1, 2, \ldots (I - 1)$ because the observed and expected proportions corrects for $k = 0$ is 0, and that for $k = I$ is 1. A collapsing algorithm combines cells until all cells have sufficiently large expected counts.

The approach of Orlando and Thissen (2000) is more intuitive than the ones using groups based on estimated abilities in that the former deals with truly observed proportion correct scores. Performing an item fit test for these models requires collapsing the original responses in some way (because, as discussed in e.g., Glas and Falcon, 2003, a test on the basis of the original responses will have poor properties for even tests of moderate lengths). Even though raw scores do not have the same appeal in 2PL or 3PL models as they do in Rasch models, they provide a basis for collapsing the original responses into groups so
that a traditional $\chi^2$-type test statistic can then be formed (whereas grouping on the basis of estimated $\theta$s does not lead to a traditional $\chi^2$-type test, as discussed earlier). Further, in the approach of Orlando and Thissen (2000), there is a naturally defined number of groups—each possible raw score defines a group. This approach should work with nonlinear tests as well, unless the lengths or overall difficulty levels of the test vary a great deal over the examinees (one way to get around the problem in these situations may be to perform the statistical test on examinees receiving the same test forms or tests of similar test length/difficulty); this area needs further investigation though.

Glas and Falcon (2003) use $S_i$ examinee groups based on raw scores for item $i$. Suppose examinee $j$ belongs to group $m_j^{(i)}$ ($m_j^{(i)} = 1, 2, \ldots S_i$) based on his/her response to items $1, 2, \ldots (i-1), (i+1), \ldots I$. The test statistic (6), for the 3PL model, for example, is based on the Lagrange multiplier test (or equivalent efficient score test) of the null hypothesis $H_0$: the 3PL model is correct vs.

$$H_1: P(y_{ij} | \theta_j, \eta_i, m_j^{(i)} = s) = c_i + (1 - c_i) \logit^{-1}(a_i(\theta_j - b_i - \beta_{is})),$$  

where $\beta_{is}$, $s = 1, 2, \ldots S_i$, gauge the deviation from the item difficulty parameter $b_i$ for the score groups. However, even though the test appears elegant, the choice of the number of groups of examinees is not obvious in the test and the model (7) may not capture all types of departures from a 3PL model. More importantly, Glas and Falcon (2003) find the overall characteristics of their test statistic worse than that suggested by Orlando and Thissen (2000); specifically, the false alarm rate for the former may be unusually high on occasions.

All of the above indicate that the test statistics suggested by Orlando and Thissen (2000) are the most acceptable item fit test as of now. Chernoff and Lehmann (1953) show that a $\chi^2$ test statistic computed from number of individuals falling into specified cells does not have a limiting $\chi^2$-distribution when estimates of parameters from the original observations are used. Instead, such a statistic is stochastically larger than what is obtained under the $\chi^2$ theory (the $\chi^2$ approximation holds however for estimates computed from the cell frequencies). The departure may be significant for small number of cells. The result then implies that the test statistics suggested by Orlando and Thissen (2000) do not have limiting $\chi^2$-distributions, because they group the examinees according to the raw scores,
but use MMLEs of item parameters from the original (ungrouped) observations. The
departure of the test statistics from the assumed $\chi^2$ distribution may be severe for a small
number of items, and there exists no recommendation as to when their $\chi^2$ approximation
gives acceptable results. Further, even a valid $\chi^2$ statistic would not perform well for small
number of examinees (because of a small number of observations in the cells) and the same
phenomenon applies to the test statistics of Orlando and Thissen (2000). Simulations in
Orlando and Thissen (2000, 2003) and Glas and Falcon (2003) show that the Type I error
rate and the false alarm rate (proportion of items that are flagged incorrectly as misfitting)
of the test statistics can be undesirably high, especially for short tests.

3. Posterior Predictive Model Checking Techniques

Guttman (1967) suggested the idea behind the posterior predictive distribution; he
used the terminology density of a future observation to describe the concept. Rubin (1981)
applied the idea to formulate the PPMC method and Rubin (1984) gave a formal Bayesian
definition of the technique. Gelman, Meng, and Stern (1996) extend the PPMC method
to allow for more direct assessment of the discrepancy between the data and the posited
model.

Let $p(y|\omega)$ denote the likelihood distribution for a statistical model applied to data
(examinee responses in this context) $y$, where $\omega$ denote all the parameters in the model.
Let $p(\omega)$ be the prior distribution on the parameters. Then the posterior distribution of
$\omega$ is $p(\omega|y) \equiv \frac{p(y|\omega)p(\omega)}{\int p(y|\omega)p(\omega)d\omega}$. Let $y^{rep}$ denote replicate data that one might observe if the
process that generated the data $y$ is replicated with the same value of $\omega$ that generated the
observed data.

The PPMC method suggests checking a model using the posterior predictive
distribution (or the predictive distribution of replicated data conditional on the observed
data),

$$p(y^{rep}|y) = \int p(y^{rep}|\omega)p(\omega|y)d\omega,$$

as a reference distribution for the observed data $y$. Basically then, the PPMC method
suggests computing an average of the likelihood distribution with respect to the posterior

8
distribution of the parameters to get the reference distribution for the data.

The next step in the PPMC method is to compare the observed data \( \mathbf{y} \) to its reference distribution (8). In practice, test quantities or discrepancy measures \( D(\mathbf{y}, \omega) \) are defined (Gelman et al., 1996), and the posterior distribution of \( D(\mathbf{y}, \omega) \) compared to the posterior predictive distribution of \( D(\mathbf{y}^{\text{rep}}, \omega) \), with any significant difference between them indicating a model failure. A researcher may use \( D(\mathbf{y}, \omega) = D(\mathbf{y}) \), a discrepancy measure depending on the data only, if appropriate.

A 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):

\[
p_b = P(D(\mathbf{y}^{\text{rep}}, \omega) \geq D(\mathbf{y}, \omega) | \mathbf{y}) = \int I_{[D(\mathbf{y}^{\text{rep}}, \omega) \geq D(\mathbf{y}, \omega)]} p(\mathbf{y}^{\text{rep}} | \omega) p(\omega | \mathbf{y}) d\mathbf{y}^{\text{rep}} d\omega,
\]

where \( I_{[A]} \) denotes the indicator function for the event \( A \).

Because of the difficulty in dealing with (8) or (9) analytically for all but simple problems, Rubin (1984) suggests simulating replicate data sets from the posterior predictive distribution in practical applications of the PPMC method. One draws \( N \) simulations \( \omega^1, \omega^2, \ldots, \omega^N \) from the posterior distribution \( p(\omega | \mathbf{y}) \) of \( \omega \), and draws \( \mathbf{y}^{\text{rep}, n} \) from the likelihood distribution \( p(\mathbf{y} | \omega^n), n = 1, 2, \ldots, N \). The process results in \( N \) draws from the joint posterior distribution \( p(\mathbf{y}^{\text{rep}}, \omega | \mathbf{y}) \), and, equivalently, from \( p(\mathbf{y}^{\text{rep}} | \mathbf{y}) \). The expression (9) suggests that the posterior predictive p-value is also the expectation of \( I_{[D(\mathbf{y}^{\text{rep}}, \omega) \geq D(\mathbf{y}, \omega)]} \), where the expectation is with respect to the joint posterior distribution \( p(\mathbf{y}^{\text{rep}}, \omega | \mathbf{y}) \). As an immediate consequence, the proportion of the \( N \) replications for which \( D(\mathbf{y}^{\text{rep}, n}, \omega^n) \) exceeds \( D(\mathbf{y}, \omega^n) \) provides an estimate of the PPP-value. Extreme posterior predictive p-values (close to 0, or 1, or both, depending on the nature of the discrepancy measure) indicate model misfit.

Gelman et al. (1996) suggest that the preferable way to perform the comparison of the realized discrepancies \( D(\mathbf{y}, \omega^n) \) and the replicated/predicted discrepancies \( D(\mathbf{y}^{\text{rep}, n}, \omega^n) \) is to plot the pairs \( \{ D(\mathbf{y}, \omega^n), D(\mathbf{y}^{\text{rep}, n}, \omega^n) \}, n = 1, 2, \ldots, N \), in a scatterplot.

Empirical and theoretical studies so far suggest that PPP-values generally have
reasonable long-run frequentist properties (Gelman et al., 1996). The PPMC method combines well with the Markov chain Monte Carlo (MCMC) algorithms (Gelman, Carlin, Stern, & Rubin, 1995).

Researchers like Robins, van der Vaart, and Ventura (2000) have shown that the PPP-values are conservative (i.e., often fails to detect model misfit), even asymptotically, for some choices of discrepancy measure, such as when the discrepancy measure is not centered. However, a conservative statistical test with reasonable power is better than a liberal test (one that rejects too often), more so for item fit, because items cost a significant amount of money and it may be to the advantage of the administrators to fail to reject the hypothesis for a borderline item and retain the item for future use. However, in order to be acceptable, the PPP-values should have adequate power to reject items that are fit extremely poorly by the model. Therefore, it will be of interest to find out whether the posterior predictive checks have enough power when the lack of fit is substantial.

4. The Discrepancy Measures, Item Fit Plots, and Bayesian P-values

This paper uses the idea of Orlando and Thissen (2000) to group the examinees according to the raw scores because the raw score is a natural and intuitive quantity.

The first discrepancy examined is the number/proportion of examinees in Group $k, k = 1, 2, \ldots (I - 1)$ who answer Item $i$ correct., $i = 1, 2, \ldots I$. Denote the proportion as $p_{ik}$, following the notation used in Section 2. For each replicated data set, there is one such proportion, denoted $p_{ik}^{rep}$. For each Item $i$, comparison of the values of $p_{ik}$ and $p_{ik}^{rep}$s, $k, k = 1, 2, \ldots (I - 1)$, provides an idea regarding the fit of the item. One way to make the comparison is to compute the posterior predictive p-value (PPP-value) for each item-group combination. This paper uses a graphical approach and suggests item fit plots in the same spirit as, for example, Hambleton and Swaminathan (1985) to make the comparison.

The Item Fit Plots

Figure 1 provides examples of two such item fit plots for a 16-item test taken by 974 examinees. The horizontal axis of the plot denotes the groups (i.e., the raw scores) of
Figure 1: Example of item fit plots.

The vertical axis represents the proportion corrects for the groups. For any group, a point denotes the observed proportion correct and a box represents the distribution of the replicated proportion corrects for that group. A line joins the observed proportions for ease of viewing. The whiskers of the box stretch till the 5th and 95th percentiles of the empirical distribution and a notch near the middle of the box denotes the median. The width of the box is proportional to the square root of the observed number of examinees in the group (the observed number in a group provides some idea about the severity of a difference between the observed and replicated proportions; a significant difference for a large group is more severe than that for a small group). For any item, too many observed proportions lying far from the center of the replicated values or lying outside the range spanned by the whiskers (i.e., lying outside a 90% prediction interval) indicate a failure of the model to explain the responses to the item. In Figure 1, the first plot (marked “Item 2”) provides such an example as two observed proportions (out of a total of 15) lie outside the 90% prediction interval and a few others lie far from the center of the box. The second plot (marked “Item 14”) shows an item that the model explains adequately, the observed values lying mostly close to the center of the boxes (always in the 90% prediction interval).

Other than providing an overall idea about the fit for an item, the suggested item fit plot also provides some idea about the region where the misfit occurs (e.g., for low or...
high-scoring individuals), the direction in which the misfit occurs (e.g., whether the model overestimates/underestimates the performance for the discrepant regions), the aspects of the item like its difficulty and slope (same type of information furnished by the item characteristic curves). One can create the plots with 95% prediction intervals as well.

The $\chi^2$-type Discrepancy Measures and Bayesian P-values

The above mentioned item-fit plots provide useful feedback about the fit of an item, but it will be useful to summarize the fit information for each item into one number, preferably a p-value. To achieve that, this paper uses the two test statistics ($\chi^2$-type and $G^2$-type) suggested by Orlando and Thissen (2000) as (realized) discrepancy measures. The $\chi^2$-type measure, denoted henceforth as $D_i^\chi(y, \omega)$, is given by

$$D_i^\chi(y, \omega) = \sum_{k=1}^{I-1} N_k \frac{(p_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})},$$

using the same notations used in (4).

The $G^2$-type measure, denoted henceforth as $D_i^G(y, \omega)$, is given by

$$D_i^G(y, \omega) = 2 \sum_{k=1}^{I-1} N_k \left[ p_{ik} \log \left( \frac{p_{ik}}{E_{ik}} \right) + (1 - p_{ik}) \log \left( \frac{1 - p_{ik}}{(1 - E_{ik})} \right) \right].$$

Each of these two statistics summarizes the fit information for an item, with large values indicating poor fit. A comparison between the posterior distribution of $D_i^\chi(y, \omega)$ [or $D_i^G(y, \omega)$] and the posterior predictive distribution of $D_i^\chi(y^{rep}, \omega)$ [$D_i^G(y^{rep}, \omega)$] provides a summary regarding the fit of Item $i$. The comparison can be done using a graphical plot (e.g., Figure 3).

Another convenient summary is the posterior predictive p-value for the discrepancy measure. Let $\omega_1, \omega_2, \ldots, \omega_N$ be a sample from the posterior distribution $p(\omega | y)$ and let $y^{rep,1}, y^{rep,2}, \ldots, y^{rep,N}$ be replicated data sets drawn from the posterior predictive distribution. The proportion of times $D_i^\chi(y^{rep,n}, \omega_n)$ [$D_i^G(y^{rep,n}, \omega_n)$] exceeds $D_i^\chi(y, \omega_n)$ [$D_i^G(y, \omega_n)$] provides an estimate of the corresponding PPP-value. Here, a PPP-value very close to 0 indicates a problem (indicating that the variability in the data set appears unusually large compared to that predicted by the model). For example, the
PPP-values corresponding to Figure 1 are 0.02 (0.01) for Item 2 and 0.68 (0.61) for Item 14, supporting the findings from the plots that the model explains the responses for Item 14 adequately while failing to do so for Item 2.

The computation of these PPP-values does not need any approximation (e.g., a $\chi^2$ approximation, as in all the existing item fit measures), and hence results in valid p-values summarizing the fit of the items. The cells (item-group combination) with small frequencies (especially for low or high raw scores) is not a problem with the PPMC approach because this does not need a $\chi^2$ assumption. However, for more stability of the discrepancy measures and the PPP-values, raw scores with too few number of examinees getting that raw score (usually high or low) can be pooled, as suggested by Orlando and Thissen (2000). Examining the distribution of the values of $D_i^\chi(y^{rep,n}, \omega_n)$ is a way to check for such stability. These quantities should look like draws from a $\chi^2$-distribution with $(I - 1 - m)$ d.f., where $I$ is the number of items in the test and $m$ is the adjustment in the degrees of freedom (d.f.) due to pooling, if there are sufficient number of examinees for each raw score—departure from the $\chi^2$ distribution will point to the instability of the discrepancy measure. This paper examines the mean, variance, and histogram of the quantities from a preliminary run of the program to ensure the stability of the measure.

Adantages and Disadvantages of the Suggested Plots and P-values

Based on the above discussion, the advantages of these item fit p-values (and item fit plots) are that

1. The Bayesian p-values are based on truly observed quantities, like those in Orlando and Thissen (2000), and hence intuitively appealing.

2. They provide natural probability statements from a Bayesian point of view.

3. Computing the p-values require neither complicated theoretical derivations nor extensive simulation studies (performed to establish the properties of a number of the existing item fit p-values).

One disadvantage of these techniques is the conservativeness of the PPMC methods (e.g., Robins et al., 2000). However, as argued earlier, a conservative test is better than a
liberal test and the conservativeness of the PPMC technique may be a boon here because items are costly and it is probably beneficial to the test administrators to retain items that are borderline. Gelman et al. (1996) comment that the PPMC method is 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 and item fit is an application exactly as referred to by them; test administrators would like to discard an item only if the model fails substantially for the item. The other disadvantage is that these techniques are based on the MCMC algorithm (e.g., Gelman et al., 1995) and hence are computation intensive. The disadvantage is more than compensated though by the above mentioned advantages. Also, a standard practice by the practitioners of the MCMC algorithm is to store the posterior sample obtained while fitting a model and to use the same to learn different aspects regarding the problem in the future; in that case, the computations to obtain the item fit measures need minutes for up to moderate-sized assessments and at most a couple of hours even for large-scale assessments (an admissions test example, later in Section 7, supports the claim). Another issue, which might make some potential users uneasy, is that these techniques are Bayesian in nature and use prior distributions. However, as Sinharay and Johnson (2003) argue, prior distributions are hardly any concern here, mainly because there exists a considerable amount of prior knowledge from previous administrations of different tests.

5. Simulation Studies: First Part

Before applying the discrepancy measures to real data sets, it is important to provide an idea about the Type I error and power of them. This paper performs detailed simulations to examine the same properties (especially the power) of the suggested item fit measures for a number of combinations of the data generating model and analysis model.

Outline of the Studies

Consider a data generating model $M_g$ and an analysis model $M_a$, where each model may be one among the 1PL, 2PL, or 3PL model. The following steps briefly describe the simulation study for the model pair. Sinharay and Johnson (2003) has more details
regarding the simulations.

1. Generate 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, their range and combination being similar to those encountered in real tests. A variety of generating parameters values ensure that the findings of the simulation study are not limited and partially 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, compute the values of observed discrepancy measures and then fit the model $M_a$ using an MCMC algorithm.

   The prior distributions used are:

   $$\log(a_j) \sim \mathcal{N}(0, 1), \quad b_j \sim \mathcal{N}(0, 1), \quad \text{logit}(c_j) \sim \mathcal{N}(-1.39, 1),$$

   chosen from our experience of analyzing data from real multiple choice tests.

3. For each 100 generated data set,

   (a) For each of the draws in the final posterior sample, generate a replicated data set and compute values of the realized and predicted discrepancy measure. A Gauss-Hermite integration routine (e.g., Thisted, 1988) with 50 points is used to perform the numerical integration (for 2- and 3PL model) required to compute the expected proportion of examinees using the approach by Lord and Wingersky (1984). Increasing the number of points causes hardly any changes in the results.

   (b) Compute the posterior predictive p-values (PPP-value) for the discrepancy by comparing the realized and predicted discrepancy measure.

   The process results in a PPP-value for each discrepancy measure for each of the 100 generated data sets.

4. Summarize, both graphically and numerically, the output of the previous program.
Results When the Model is Fitted is the Data Generating Model

These analyses should provide ideas about the Type I error rate of the suggested item fit diagnostics. Vertical lines in Figure 2 show the distributions for the 100 p-values for the $\chi^2$-type discrepancy measure for all the items when 3PL model is fitted to data generated from the 3PL model. The lines extend till the 5th and 95th percentiles of the PPP-values, while an “x” towards the middle of the line denotes the median PPP-value. There is a horizontal line at 0.05—a p-value lying below this line indicates a problem with the model. The p-values mostly concentrate around 0.5 and the lower ends of the lines rarely extend below 0.05. The overall proportion of times the PPP-values are extreme at 5% level (and hence indicate a problem with the model) is 0.04. The same is true for the $G^2$-type test statistic as well, which is different from the findings in Orlando and Thissen (2000), who find the measure to have higher Type I error (maybe because they used 1,000 examinees while the number is 2,500 for this study). These results show that the Bayesian item fit measures do not indicate a problem (with a model) when it is not supposed to do so. The measure performs very similarly for data generated from the Rasch model or 2PL model (output not shown) and analyzed using the same model.

![Figure 2: Distribution of the item fit p-values when the 3PL model is fit to 3PL data.](image)

Results When Model Fitted is Different from the Data Generating Model

These analyses should provide an idea of the power of the suggested p-values.
**Fit of the Rasch Model to 2PL/3PL Data**

Figure 3 shows, for two items, plots of the realized discrepancy $D_i^X(y, \omega)$ and the predicted discrepancy $D_i^X(y^{rep}, \omega)$ when the Rasch model is fitted to one data set generated from the 3PL model. A diagonal line is there for convenience—points consistently below the diagonal (showing that the realized discrepancy is mostly larger than the predicted discrepancy) indicate a problem. The plot towards the left shows that the model cannot explain the responses for the item adequately (with a true discrimination parameter of 0.7 only) very well, the model-predicted discrepancy being almost always smaller than the realized discrepancy. The PPP-value is 0.00. The plot towards the right shows the opposite—the model seems to reproduce the discrepancy and hence explain the responses for the item (a true discrimination of 1.10) adequately (PPP-value=0.35).

![Figure 3: Comparison of realized and predicted discrepancies when the Rasch model is fit to 3PL data.](image)

Figure 4 shows distributions of the PPP-values for the $\chi^2$-type discrepancy measure for all the items when Rasch model is fitted to data generated from the 3PL model. The items are sorted by the values of the true discrimination parameters ($a_j$s) (which are shown on the axis at the bottom) for them. The plot shows a very interesting pattern. The p-values are mostly around 0.5 for items whose discrimination parameters lie within a certain range—the Rasch model can explain these items adequately so that the p-values does not have enough power for these items. This result is very similar to that in Orlando and Thissen (2000), who find their item fit statistics to have low power for true $a_j$s lying in the middle of the range. Because the Rasch model assumes a common discrimination
Figure 4: Distribution of the Item fit p-values when the Rasch model is fit to 3PL data.

Parameter for all the items in a test, the model can explain items generated from a 3PL model with true discriminations around that common value. The overall proportion of times the PPP-values are extreme at 5% level (and hence indicate a problem with the model) is 62, which is slightly higher than the corresponding proportions in Orlando and Thissen (2000). The results are similar when the Rasch model is fitted to data generated from a 2PL model and are not shown.

Figure 5 shows item fit plots for four items for the Rasch model fit to one data set from the 3PL model, along with the true discrimination parameters for the items. The p-values for the $\chi^2$-type discrepancy are 0.00, 0.01, 0.00, and 0.35 respectively; that is, the model cannot adequately explain any but the last item in the Figure. The plots show how the model fails to explain the responses for the first three of the above four items. For the first item (discrimination 0.7), the model fails mostly for middle and high score range; the model fails for the second item (discrimination 0.56) for the whole score range, appearing to fit the item very poorly; for the third item (discrimination 2.4), the model fails for all but very high and very low scores.

The behavior of the item fit plots and the test statistics are very similar when the Rasch model is fit to data from the 2PL model and are skipped.
Figure 5: Item fit plots for four items for the Rasch model fit to 3PL data.

Fit of the 2PL Model to 3PL Data

Figure 6 shows distributions for the PPP-values for the $\chi^2$-type discrepancy measure for all the items when 2PL model is fitted to data generated from the 3PL model. The items are sorted by the values of their true guessing parameters ($c_j$s, which are shown on the axis at the bottom). The p-values are mostly around 0.5 for most items, showing, as in Orlando and Thissen (2000), that the 2PL model can explain most of the items generated from the 3PL model. The overall power is 0.12, very similar to that in Orlando and Thissen
Figure 6: Item fit p-values when the 2PL model is fit to 3PL data.

(2000) again. The 2PL model cannot fit the items that have high true guessing parameter and high difficulty—there seem to be three of them here.

For all the simulation studies above, the performance of the $G^2$-type discrepancy measure is very similar to that of the $\chi^2$-type measure and hence we do not show results for the measure.

Discussion

The simulation studies show that the suggested Bayesian item fit measures have reasonable Type I error, i.e., they will not wrongly indicate a misfit too often. Whenever data come from 1-, 2-, or 3PL model and the fitted model is the same as the generating model, the proportion of times the PPP-values are extreme at 5% level never exceeds 0.05.

The studies also show that the measures have considerable power, that is, they will indicate misfit when they are supposed to do so. When the Rasch model is fitted to data from 2PL or 3pl model, the measures demonstrate adequate power to flag items that the Rasch model cannot explain. Because the Rasch model is a special case of both the 2PL model and 3PL model, the former model can explain some items generated from the latter two models adequately—the item fit measures do not indicate any problem, and rightly so, with those items.

When a 2PL model is fitted to data from 3PL model, the fit measures have little
power; Orlando and Thissen (2000), Yen (1981), and McKinley and Mills (1985) make the same observation. For most combination of parameter values, the 2PL model has the ability to adjust its parameters so that the fitted (2PL) ICC is close to the generating ICC (from a 3PL model), the only difference occurring for extreme proficiency values, where there are few individuals. As a result, the realized item fit measures remain small and the item fit p-value is not extreme. The plot in the left side of Figure 7, showing what happens when the 2PL model is fitted to one data set generated from the 3PL model, depicts the situation (PPP-value=0.63). The only items for which the 2PL model fails to do so are those with high guessing parameter and high difficulty, as exemplified by the plot in the right side of Figure 7. Despite the efforts of the 2PL model to make up for the high (0.25) guessing

![An Item 2PL model can explain](image1.png)  ![An Item 2PL model cannot explain](image2.png)

**Figure 7**: Fitted (2PL) and true (3PL) ICCs for two items when the 2PL model is fitted to 3PL data

parameter, there is still visible differences between the true ICC and the fitted ICC in the middle of the range of \( \theta \). The PPP-value for the item for this simulated data set is 0.032.

Overall, the performance of the item fit measures appear satisfactory in assessing fit of the unidimensional IRT models from the simulation studies.
6. Simulation Studies: Second Part

This paper performs another set of simulation studies, mainly to establish the properties of the Bayesian item fit p-values under different circumstances. The main objective here is to study under different situations the Type I error rate, hit rate (proportion of items flagged correctly as misfitting), and false alarm rate (proportion of items flagged incorrectly as misfitting) of the Bayesian item fit statistics. As mentioned earlier, the item fit measures suggested by Orlando and Thissen (2000, 2003) and Glas and Falcon (2003) both suffer from undesirable false alarm rates when the data set has a few truly misfitting items—it will be of interest to examine how the Bayesian item fit measures perform in those situations. Also, this section is meant to demonstrate the superiority of the Bayesian item fit measures over the others for short tests and/or tests given to small number of examinees.

**Studying the Type I Error Rate**

In one set of simulations, we generated data from the 3PL model, fitted the 3PL model to the data for different combinations of test length (10, 40, 78) and number of examinees (500, 1,000, 4,000) and studied the Type I error rates of the Bayesian item fit p-values. The same program also computed the p-values suggested by Orlando and Thissen (2000) using the posterior mean of the parameters instead of the MMLEs of them. This may provide p-values that are slightly different from those obtained with the MMLEs, but the difference will be negligible given that we used noninformative priors on the parameters, making the MMLEs and the posterior means very close to each other. Patz and Junker (1999b,a) demonstrate the closeness of MMLEs and posterior means for 2PL and 3PL models. Further, using posterior mean will remove any effect on the p-values of difference between the Bayesian estimation and maximum likelihood estimation of parameters.

The generating item parameters used were:

- For 10 items, as in Glas and Falcon (2003), the data generator used a constant \( c_i = 0.2 \), and three values of \( a_i \)s, 0.5, 1, and 1.5, crossed with three values of \( b_i \)s, -1, 0, and 1 to produce nine sets of parameters. The 10th item had the combination \( a_i = 1, b_i = \)
0, c_i = 0.2.

- For 30 items, the data generator used three values of a_i's, 0.5, 1.25, and 2, crossed with five values of b_i, -2, -1, 0, 1, and 2, and two values of c_i, 0.15, and 0.25.

- For 78 items, the data generator used MMLEs for a real 78-item multiple choice test (considered in Sinharay, 2003).

Table 1: The Type I Error Rates for the PPP-values Corresponding to the χ^2-type and G^2-type Discrepancy Measure

<table>
<thead>
<tr>
<th>No. of items</th>
<th>No. of examinees</th>
<th>Bayesian χ^2</th>
<th>Bayesian G^2</th>
<th>S − χ^2</th>
<th>S − G^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>500</td>
<td>0.03</td>
<td>0.03</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.08</td>
<td>0.07</td>
</tr>
<tr>
<td>30</td>
<td>500</td>
<td>0.03</td>
<td>0.03</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.03</td>
<td>0.04</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>0.04</td>
<td>0.03</td>
<td>0.08</td>
<td>0.04</td>
</tr>
<tr>
<td>78</td>
<td>500</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Table 1 provides the overall proportion of times the PPP-values are extreme at 5% level (i.e., the Type I error rate). The table also shows the Type I error levels of the indices (denoted as S − χ^2 and S − G^2) by Orlando and Thissen (2000). The results demonstrate that the Bayesian p-values are slightly conservative—the Type I error rate never exceeds the nominal 5% level and sometimes slightly less than the level. The indices suggested by Orlando and Thissen (2000) suffer from higher Type I error rate than what is desired, especially for short tests (even though Orlando and Thissen, 2003, recommends use of their index for short tests) and small sample sizes. The Type I error rates for S − G^2 in the above table are very close to those found in Glas and Falcon (2003) and hence probably are not outcomes of the use of posterior means instead of MMLEs.
**Studying the Power and False Alarm Rate Under Model Violation**

A second set of studies repeated the above simulations, except that while generating the data sets, the data generator mixed a low percentage (10 or 20) of misfitting/bad items along with the 3PL items. The ICCs for the bad items were one of four types:

- **Bad Item 0:** An ICC considered by Glas and Falcon (2003) and given by the right-hand side of (7), for $S_i = 5$, and $\beta_i$s 0, -0.5, 0.5, -0.5, and 0.

- **Bad Item 1:** One with ICC

  $$P(y_{ij} = 1 \mid \theta, \eta_i) \equiv 0.25 \logit^{-1} \left(-4.25 (\theta - 1 + 1.5) + \logit^{-1} (4.25 (\theta - 1))\right),$$

  that is nonmonotone for low $\theta$.

- **Bad Item 2:** One with ICC $P(y_{ij} = 1 \mid \theta, \eta_i) \equiv 0.7 \logit^{-1} (3.4 (\theta + 0.5))$, that does not go to 1 as $\theta \to \infty$.

- **Bad Item 3:** One whose ICC is a mixture of two logistic functions,

  $$P(y_{ij} = 1 \mid \theta, \eta_i) \equiv 0.55 \logit^{-1} (5.95 (\theta + 1)) + 0.45 \logit^{-1} (5.95 (\theta + 1 - 3)).$$

  This ICC exhibits a plateau over middle values of $\theta$, but follows a logistic curve before and after a plateau.

Orlando and Thissen (2003) draw the ICCs for Bad Items 1-3 (with the corresponding best-fitting 3PL ICC), which they find to have occurred in real educational assessments.

Glas and Falcon (2003) and Orlando and Thissen (2003) observe that when the test has a low percentage of misfitting items, their item fit measures have occasional undesirably high false alarm rates (i.e., the proportion of items incorrectly flagged as misfitting). For example, Table 4 in Glas and Falcon (2003) shows that for a test with 10 items and 4,000 examinees, if one item is Bad Item 0 (other nine from a 3PL model), the false alarm rate for the index by Glas and Falcon (2003) is 0.92, while that for $S - G^2$ is 0.20.

Tables 2 and 3 provide the overall proportion of times the PPP-values are extreme at 5% level for the misfitting item (the power) and the same for the 3PL items (false alarm rate) for the Bayesian measure. The tables also show the same quantities for indices
(denoted as $S - \chi^2$ and $S - G^2$) by Orlando and Thissen (2000). The type of Bad Items used in the simulations for each case are shown as well.

Table 2: The Power and False Alarm Rate for the PPP-values Corresponding to the $\chi^2$-type and $G^2$-type Discrepancy Measure for 10% Misfitting Items

<table>
<thead>
<tr>
<th>No. of items</th>
<th>No. of examinees</th>
<th>Bayesian $\chi^2$</th>
<th>Bayesian $G^2$</th>
<th>$S - \chi^2$</th>
<th>$S - G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>10/1,000</td>
<td>0.28</td>
<td>0.71</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td>Item 0</td>
<td>4,000</td>
<td>1.00</td>
<td>0.74</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>10/1,000</td>
<td>0.11</td>
<td>0.11</td>
<td>0.26</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td>4,000</td>
<td>0.93</td>
<td>0.93</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td>30/1,000</td>
<td>0.31</td>
<td>0.31</td>
<td>0.40</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>4,000</td>
<td>0.72</td>
<td>0.63</td>
<td>0.81</td>
<td>0.70</td>
</tr>
<tr>
<td>78/1,000</td>
<td>0.47</td>
<td>0.30</td>
<td>0.53</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3</td>
<td>4,000</td>
<td>0.88</td>
<td>0.83</td>
<td>0.91</td>
<td>0.74</td>
</tr>
<tr>
<td>False alarm rate</td>
<td>10/1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.14</td>
<td>0.11</td>
</tr>
<tr>
<td>Item 0</td>
<td>4,000</td>
<td>0.05</td>
<td>0.04</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>10/1,000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Item 3</td>
<td>4,000</td>
<td>0.04</td>
<td>0.03</td>
<td>0.13</td>
<td>0.11</td>
</tr>
<tr>
<td>30/1,000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.09</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>4,000</td>
<td>0.04</td>
<td>0.03</td>
<td>0.07</td>
<td>0.05</td>
</tr>
<tr>
<td>78/1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.07</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3</td>
<td>4,000</td>
<td>0.04</td>
<td>0.05</td>
<td>0.06</td>
<td>0.04</td>
</tr>
</tbody>
</table>

The tables reveal a number of interesting facts:

1. The false alarm rate of the Bayesian item fit measures is never larger than the Type I error level. Remembering the high false alarm rates of the existing item fit measures...
Table 3: The Power and False Alarm Rate for the PPP-values Corresponding to the $\chi^2$-type and $G^2$-type Discrepancy Measure for 20% Misfitting Items

<table>
<thead>
<tr>
<th>No. of items</th>
<th>No. of Item-type</th>
<th>No. of examinees</th>
<th>Bayesian $\chi^2$</th>
<th>Bayesian $G^2$</th>
<th>$S - \chi^2$</th>
<th>$S - G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power</td>
<td>10</td>
<td>500</td>
<td>0.10</td>
<td>0.12</td>
<td>0.30</td>
<td>0.28</td>
</tr>
<tr>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.23</td>
<td>0.26</td>
<td>0.42</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>2, 3</td>
<td>4,000</td>
<td>0.91</td>
<td>0.90</td>
<td>0.96</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>500</td>
<td>0.70</td>
<td>0.58</td>
<td>0.80</td>
<td>0.63</td>
</tr>
<tr>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.72</td>
<td>0.65</td>
<td>0.83</td>
<td>0.69</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>4,000</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>500</td>
<td>0.40</td>
<td>0.31</td>
<td>0.45</td>
<td>0.22</td>
</tr>
<tr>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.86</td>
<td>0.79</td>
<td>0.89</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3</td>
<td>4,000</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>False alarm</td>
<td>10</td>
<td>500</td>
<td>0.04</td>
<td>0.04</td>
<td>0.15</td>
<td>0.14</td>
</tr>
<tr>
<td>rate</td>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>2, 3</td>
<td>4,000</td>
<td>0.05</td>
<td>0.05</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>500</td>
<td>0.03</td>
<td>0.02</td>
<td>0.08</td>
<td>0.05</td>
</tr>
<tr>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.04</td>
<td>0.05</td>
<td>0.10</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>4,000</td>
<td>0.04</td>
<td>0.03</td>
<td>0.09</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>500</td>
<td>0.03</td>
<td>0.02</td>
<td>0.06</td>
<td>0.02</td>
</tr>
<tr>
<td>/Bad Items</td>
<td>1,000</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>0, 1, 2, 3</td>
<td>4,000</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

found by Glas and Falcon (2003) and Orlando and Thissen (2003), and looking at the false alarm rates for the measures $S - \chi^2$ and $S - G^2$ in the above tables, the Bayesian item fit measure performs far better than those item fit measures. A good statistical test should not reject the null hypothesis incorrectly too often and the Bayesian item fit measures satisfy that requirement.

2. The power for the indices suggested by Orlando and Thissen (2000) is always more than that for the Bayesian item fit measures. The difference is significant for 500 or 1,000 examinees, while not so (as the power is close to 1 for all the measures) for large number (4,000) of examinees. Combined with the false alarm rate, this fact shows that the Bayesian tests are conservative in comparison to the tests in Orlando and
Thissen (2000). The former may have low power for small sample size for some types of bad items, implying that the test is reluctant to flag an item as misfitting unless there is enough evidence against it. The conservativeness of the Bayesian p-values is a combination of the conservativeness of the PPMC method and the ability of the 3PL model to adequately explain items with ICCs that are slightly different from the logistic form (a phenomenon observed by, e.g., Orlando and Thissen, 2003).

3. Individual bad items are differently “bad” regarding how the 3PL model can fit them. For example, the 3PL model succeeds in estimating the ICC for Bad Item 3 pretty well for a small number of examinees so that the power of the tests to detect misfit for that item is low (same result found as in Orlando and Thissen, 2003). On the other hand, the power is much more for the Bad Item 0.

The above simulation results suggest that the Bayesian item fit measures are superior to the existing item fit measures. The Bayesian p-values have low Type I error rates and false alarm rates (unlike any of the existing item fit p-values)—they mostly do not indicate a problem with an item when they are not supposed to. They have low power for small sample size, but rightfully so—the basic principle of testing of hypothesis is that one should not reject the null unless there is sufficient reason to do so. The measures have considerable power for even moderately large sample size, that is, when there is substantial evidence against the model. Use of this measure will lead to rejection of borderline items less often, which can lead to significant savings because test items, depending on complexity, may be quite expensive, and test administrators would like to discard an item only when absolutely justified.

**Item Fit Plots**

Figure 8 shows the item fit plots for one data set each for each of the four types of bad items considered in the above simulation studies. The PPP-value is zero for each of these four situations. The plot for Bad Item 0 is for the case with 10 items and 4,000 examinees; that for Bad Items 1-3 are for the case with 30 items and 4,000 examinees. The plots seem to demonstrate the type of violations from the 3PL model quite well. For
Figure 8: Item fit plots corresponding to the four types of bad items.

example, the plot for Bad Item 1 shows the nonmonotone nature of the observed proportion corrects for raw scores less that 20—the predicted values are far apart from the observed values. Similarly, the plot for the Bad Item 3 shows the fluctuating nature of the difference of the observed and predicted proportion corrects.

7. Real Data Examples

Mixed Number Subtraction

Data

Let us consider a part of the often analyzed mixed number data set from Tatsuoka (1994). The data set considered has responses of 325 middle-school students to 15 mixed
number subtraction problems for which the fractions do have a common denominator. The items were designed to diagnose student mastery of five skills: (a) basic fraction subtraction, (b) simplify/reduce fraction or mixed number, (c) separate whole number from fraction, (d) borrow one from the whole number in a given mixed number, and (e) convert a whole number to a fraction.

Table 4 lists the 15 items considered, characterized by the skills they require. Horizontal lines in the table separate the items in groups according to the skill requirements. A number of papers, starting from Mislevy (1995), analyze the data set using a Bayesian network model. The data set is small so that the asymptotic $\chi^2$-approximations of the test statistics in Orlando and Thissen (2000) are likely to fail.

Table 4: Skill Requirements for the Mixed Number Subtraction Problems

<table>
<thead>
<tr>
<th>Item No</th>
<th>Text of the item</th>
<th>Skills required</th>
<th>Proportion correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 4</td>
<td>$\frac{5}{7} - \frac{1}{7}$</td>
<td>x</td>
<td>0.79</td>
</tr>
<tr>
<td>4 8</td>
<td>$\frac{3}{4} - \frac{3}{4}$</td>
<td>x</td>
<td>0.70</td>
</tr>
<tr>
<td>9 11</td>
<td>$\frac{3}{4} - \frac{3}{4}$</td>
<td>x</td>
<td>0.75</td>
</tr>
<tr>
<td>10</td>
<td>$\frac{1}{8} - \frac{1}{8}$</td>
<td>x</td>
<td>0.71</td>
</tr>
<tr>
<td>11</td>
<td>$\frac{3}{5} - 2$</td>
<td>x</td>
<td>0.69</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{3}{5} - \frac{3}{5}$</td>
<td>x</td>
<td>0.37</td>
</tr>
<tr>
<td>7 12</td>
<td>$\frac{4}{3} - \frac{4}{3}$</td>
<td>x</td>
<td>0.37</td>
</tr>
<tr>
<td>12</td>
<td>$\frac{7}{5} - \frac{4}{5}$</td>
<td>x</td>
<td>0.34</td>
</tr>
<tr>
<td>13</td>
<td>$\frac{4}{10} - \frac{2}{10}$</td>
<td>x</td>
<td>0.31</td>
</tr>
<tr>
<td>10</td>
<td>$2 - \frac{1}{8}$</td>
<td>x</td>
<td>0.38</td>
</tr>
<tr>
<td>3 14</td>
<td>$3 - \frac{2}{5}$</td>
<td>x</td>
<td>0.33</td>
</tr>
<tr>
<td>14</td>
<td>$7 - \frac{1}{4}$</td>
<td>x</td>
<td>0.26</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{4}{12} - \frac{2}{12}$</td>
<td>x</td>
<td>0.31</td>
</tr>
</tbody>
</table>
Analysis

We fit the 1-, 2-, and 3PL models to the data set using an MCMC algorithm (Sinha, 2003). The prior distributions used are

\[
\log(a_j) \sim \mathcal{N}(0, 1), \quad b_j \sim \mathcal{N}(0, 1), \quad \text{logit}(c_j) \sim \mathcal{N}(-2.2, 1),
\]

where, \( \text{logit}(0.1) = -2.2 \). Because the items are not multiple choice, the a priori probability for getting them correct just by guessing is assumed to be around 0.1. We then create item fit plots and compute the PPP-values for the discrepancies for the above mentioned models. For comparison purposes, we also compute the classical p-values suggested by Orlando and Thissen (2000) for the 3PL model under a \( \chi^2 \)-assumption (and MMLEs of the item parameters obtained using PARSSCALE, Muraki and Bock, 1991).

Item Fit Plots

Figure 9 shows the item fit plots for the Rasch model fit to the data set. The plot shows that the Rasch model does a poor job regarding item fit. There is considerable overall difference between the observed and replicated proportion-corrects for most of the items, with the most visible difference occurring for Items 3, 5, and 10.

The situation is better for the 3PL model. Figure 10 shows the item fit plots for the model. Still, there are some items, such as Items 3 and 5, for which the observed proportion-corrects differ substantially from the predicted proportions. Plots for both these items (for the 3PL model) have some similarity to that for “Bad Item 0” and “Bad Item 3” in Figure 8.

These plots provide useful feedback. Other than showing if the model explains the items well or not, they also show the directions of departure of the predicted proportion-corrects from the corresponding observed values. For example, the plot for Item 5 in Figure 9 shows that the model underestimates the proportion-corrects for the item for examinees with low raw scores, but the pattern changes for examinees with high raw scores. The plot points to the inflexibility of the Rasch model to adjust to the observed proportions. Figure 10 shows that the fit of the 3PL model to Item 5 is better—the model can adapt to the data better because it has a guessing parameter.
Figure 9: Item fit plots corresponding to the Rasch model fit to the mixed number data set.

P-values

The item fit plots above are informative, but practitioners will find a measure that quantifies the fit information for each item even more useful. The posterior predictive
Figure 10: Item fit plots corresponding to the 3PL model fit to the mixed number data set.

p-value is one such measure. Table 5 shows the posterior predictive p-values (rounded to two decimal places) for the Bayesian item fit measures (10) and (11) for each item when the data are analyzed using the 1-, 2-, and 3PL model. For the 3PL model, the table also
Table 5: The Bayesian P-values for 1-, 2-, and 3PL Model and Classical P-values for the 3PL Model Fitted to the Mixed Number Subtraction Data

<table>
<thead>
<tr>
<th>Item no.</th>
<th>3PL-classical $\chi^2$</th>
<th>3PL-Bayesian $\chi^2$</th>
<th>3PL-Bayesian $G^2$</th>
<th>2PL-Bayesian $\chi^2$</th>
<th>2PL-Bayesian $G^2$</th>
<th>1PL-Bayesian $\chi^2$</th>
<th>1PL-Bayesian $G^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.58 0.64</td>
<td>0.35 0.24</td>
<td>0.57 0.37</td>
<td>0.02* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.01* 0.05*</td>
<td>0.02* 0.00*</td>
<td>0.02* 0.01*</td>
<td>0.00* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.00* 0.01*</td>
<td>0.01* 0.01*</td>
<td>0.01* 0.00*</td>
<td>0.00* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.00* 0.00*</td>
<td>0.24 0.20</td>
<td>0.03* 0.02*</td>
<td>0.04* 0.04*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.01* 0.02*</td>
<td>0.05 0.07</td>
<td>0.22 0.23</td>
<td>0.00* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.45 0.55</td>
<td>0.16 0.13</td>
<td>0.25 0.17</td>
<td>0.08 0.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.53 0.58</td>
<td>0.55 0.42</td>
<td>0.00* 0.00*</td>
<td>0.05 0.01*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.01* 0.03*</td>
<td>0.02* 0.01*</td>
<td>0.01* 0.01*</td>
<td>0.00* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.07 0.06</td>
<td>0.08 0.10</td>
<td>0.11 0.05*</td>
<td>0.01* 0.00*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.04* 0.02*</td>
<td>0.06 0.08</td>
<td>0.07 0.04*</td>
<td>0.01* 0.01*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>0.43 0.52</td>
<td>0.30 0.12</td>
<td>0.45 0.21</td>
<td>0.08 0.01*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>0.40 0.54</td>
<td>0.44 0.39</td>
<td>0.30 0.27</td>
<td>0.49 0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0.33 0.33</td>
<td>0.22 0.19</td>
<td>0.28 0.22</td>
<td>0.12 0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>0.57 0.76</td>
<td>0.65 0.81</td>
<td>0.75 0.70</td>
<td>0.16 0.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>0.79 0.98</td>
<td>0.64 0.74</td>
<td>0.19 0.50</td>
<td>0.18 0.34</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. An asterisk indicates a p-value extreme at 5% level.*

shows the classical p-values corresponding to the test statistics in Orlando and Thissen (2000). For convenience, the p-values extreme at 5% level are marked with an asterisk. For computing the classical item fit p-values, we pool the Raw Scores 1 and 2 into one group because the expected counts for Raw Score 1 is less than one for a few items; the same is true for Raw Scores 13 and 14. To make the comparison meaningful, we use similar pooling for computing the Bayesian p-value as well.

It is important to note that the Bayesian p-values are significant for a number of items even for such a small data set, corroborating the findings from the simulations that the PPP-values may have considerable power to detect item misfit even for small data sets. A comparison of the classical and Bayesian p-values (for the 3PL model) suggests some similarities as well as some differences. Both sets of p-values indicate that the model is not adequate for Items 2, 3, and 8, whereas both sets agree that the model is adequate for Items 1, 6, 7, 9, 11, 12, 13, 14, and 15 (there are some differences about the magnitude of p-values though).
The two methods differ for the remaining items. The most glaring difference occurs for Item 4. Both the classical p-values are extreme whereas the Bayesian p-values are 0.2 and 0.24. The plot for the item shows that although there is a considerable difference between the observed and predicted proportion for Raw Score 2, the model does not perform too badly for the middle of the score range, where lie most of the examinees; therefore, the Bayesian p-values for the item not being significant is probably justified.

Because of the rather small number (15) of items, the $\chi^2$ approximation of the statistics of Orlando and Thissen (2000) is probably not appropriate here; the true null distribution is (significantly) stochastically larger than the $\chi^2$ distribution assumed, because of the effect mentioned by Chernoff and Lehmann (1953). Further, a small number of examinees (325), causing a number of small cell counts, adds to the problem. Therefore, the statistics have Type-I error levels much higher than the intended 5%. There are differences in the Bayesian p-values and frequentist p-values for Items 5 and 10 as well, although the differences are much smaller for these items.

The difference in the Bayesian and frequentist p-values are not outcomes of the prior distributions. PPP-values under other prior distributions (e.g., those with lower prior means for $c$'s, remembering that the items are not multiple choice) are slightly different, but their significance status does not change as long as noninformative prior distributions are used, so that the difference of the PPP-values from those suggested by Orlando and Thissen (2000) cannot be explained by the prior distributions.

The performance of the 2PL model does not appear to be much different from that of the 3PL model so far as item fit is concerned, although there are a few more extreme p-values for the 2PL model. Even though the items are not multiple choice, the 3PL model can explain responses to Items 4 and 7 reasonably well, where the 2PL model fails to do so. The Rasch (1PL) model performs considerably worse than the 3PL (and 2PL) model, with 8 and 10 extreme p-values for the $\chi^2$-type measure and the $G^2$-type measure respectively. Overall, none of the models seem to perform too satisfactorily for the data set regarding item fit. Analysis in Sinharay (2003) shows that the models are not adequate in explaining a number of other aspects of the data set either.
An Example From the National Assessment of Educational Progress

Consider a subsample from the National Assessment of Educational Progress (NAEP) Math OnLine (MOL) special study (e.g., Johnson and Sinharay, 2003) that consists of the responses of 974 8th grade examinees to 16 multiple choice items in one test form. We fit the 1-, 2-, and 3PL models to the data set using an MCMC algorithm. The prior distributions used are the same as those in (12) except that the prior distribution on the guessing parameters is set as $\text{logit}(c_j) \sim \mathcal{N}(-1.4, 1)$ because there are five options for each item and $\text{logit}(1/5) = -1.39$. We then create the item fit plots and compute the PPP-values for the discrepancies. The classical p-values for the statistics by Orlando and Thissen (2000) are also computed.

**Item Fit Plots**

Figure 11 shows the item fit plots for the 3PL model fit to the data set. The plot shows that the 3PL model does well overall regarding item fit. There is considerable overall difference between the observed and replicated proportion corrects for Items 2 and 8 only.

**P-values**

For computing the classical item fit p-values, we pool the Raw Scores 1-2 into one group because the expected count for that raw score is less than 1 for a few items; the same is true for Raw Scores 14-15. To make the comparison meaningful, we use similar pooling for computing the Bayesian p-values as well. Because the p-values for the $\chi^2$-type and the $G^2$-type statistic are very similar, we only describe the results with the $\chi^2$-type measure for this data set.

Table 6 shows the posterior predictive p-values (rounded to two decimal places) for each item when the data are analyzed using the 1-, 2-, and 3PL model. For the 3PL model, the table also shows the classical p-values corresponding to the test statistics in Orlando and Thissen (2000). For convenience, the p-values extreme at 5% level are marked with an asterisk.

For the 3PL model, the classical and Bayesian p-values are in good agreement,
Figure 11: Item fit plots corresponding to the 3PL model fit to the NAEP data set.

which probably shows that the $\chi^2$ approximation in Orlando and Thissen (2000) is more accurate than that in the previous example, may be because of a larger number of examinees (even though the number of items here is 16, only one more than that in the previous example).
Table 6: The Bayesian P-values for 1-, 2-, and 3PL Model and Classical P-values for the 3PL Model Fitted to the NAEP Data

<table>
<thead>
<tr>
<th>Item no.</th>
<th>3PL-classical</th>
<th>3PL-Bayesian</th>
<th>2PL-Bayesian</th>
<th>1PL-Bayesian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$G^2$</td>
<td>$\chi^2$</td>
<td>$G^2$</td>
</tr>
<tr>
<td>1</td>
<td>0.20</td>
<td>0.20</td>
<td>0.33</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>0.01*</td>
<td>0.01*</td>
<td>0.02*</td>
<td>0.01*</td>
</tr>
<tr>
<td>3</td>
<td>0.32</td>
<td>0.30</td>
<td>0.42</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>0.87</td>
<td>0.86</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td>5</td>
<td>0.24</td>
<td>0.31</td>
<td>0.27</td>
<td>0.30</td>
</tr>
<tr>
<td>6</td>
<td>0.20</td>
<td>0.24</td>
<td>0.38</td>
<td>0.41</td>
</tr>
<tr>
<td>7</td>
<td>0.46</td>
<td>0.76</td>
<td>0.52</td>
<td>0.65</td>
</tr>
<tr>
<td>8</td>
<td>0.13</td>
<td>0.15</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>9</td>
<td>0.94</td>
<td>0.94</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>10</td>
<td>0.41</td>
<td>0.77</td>
<td>0.39</td>
<td>0.66</td>
</tr>
<tr>
<td>11</td>
<td>0.54</td>
<td>0.62</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>12</td>
<td>0.89</td>
<td>0.88</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>13</td>
<td>0.31</td>
<td>0.27</td>
<td>0.40</td>
<td>0.33</td>
</tr>
<tr>
<td>14</td>
<td>0.93</td>
<td>0.93</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>15</td>
<td>0.60</td>
<td>0.57</td>
<td>0.68</td>
<td>0.61</td>
</tr>
<tr>
<td>16</td>
<td>0.35</td>
<td>0.33</td>
<td>0.45</td>
<td>0.43</td>
</tr>
</tbody>
</table>

*Note.* An asterisk indicates a p-value extreme at 5% level.

*An Admissions Test Example*

This example relates to the responses of a random subset of 10,000 examinees to 78 multiple choice items of the verbal measure of a widely used admissions test. The objective in this example is to find out how the standard IRT models perform, regarding item fit, in a real large-scale assessment. As this paper deals with dichotomous items, we treat the omitted and not reached responses as wrong answers.

The item types, in the order in which they appear in the test booklet received by the subsample of examinees considered, are as follows:

1. Items 1-10 and 37-45: discrete/stand-alone items requiring an examinee to complete a sentence (“completion items” henceforth)

2. Items 11-23 and 46-51: discrete items involving word relationships (“WR items” for future reference)
3. Items 24-36, 51-57, 58-66, 67-78: items based on a shared stimulus material (like a reading passage)

Items 1-36 form the first separately timed section. Items 37-66 form the second section and Items 67-78 form the third section. There is a strict time limit for each of the three sections, and evidence suggests that the test may be slightly speeded.

**Analysis**

We fit the 1-, 2-, and 3PL models to the data set using an MCMC algorithm. The prior distributions used are the same as those in (12) except that the prior distribution on the guessing parameters is set as $\logit(c_j) \overset{iid}{\sim} \mathcal{N}(-1.4, 1)$ because there are five options for each item and $\logit(1/5) = -1.39$.

We then compute the PPP-values for the discrepancies and create item fit plot for a number of items. Even though the data set is large ($10,000 \times 78$ responses), the whole process takes about 12 hours for the 3PL model on a Sun Blade 1000 workstation equipped with a 950 MHz CPU and a half gigabyte of RAM. However, because the MCMC algorithm was run and its output stored as a part of the computations in Sinharay (2003), it took about an hour to perform the computations for the item fit analysis.

**P-values**

For computing the classical item fit p-values, we pool the Raw Scores 1-12 into one group because the expected count for that raw score is less than one for a few items; the same is true for Raw Scores 74-77. To make the comparison meaningful, we use the same pooling for computing the Bayesian p-values as well. Because the p-values for the $\chi^2$-type and the $G^2$-type statistic are very similar, we only describe the results with the $\chi^2$-type measure for this data set.

Figure 12 plots the classical item fit p-values and the Bayesian p-values for the 3PL model fit to the data set. The plot shows horizontal and vertical lines at 0.05 and a diagonal line for convenience. The two sets of p-values agree quite well, which is not surprising given the large number of items and examinees for the data set. The only item for which the
Figure 12: Posterior predictive p-values vs. the classical p-values for the 3PL model fit to the admissions test data set.

there is a difference in the conclusions of the two methods is Item 48, for which the classical p-value is Bayesian p-value that is 0.04 while the Bayesian p-value is 0.054, which is a minor difference. For all the other 77 items, either both p-values are significant at 5% level (15 such items, spread uniformly among the three item types, indicating that the model does not seem to be adequate regarding item fit) or both are nonsignificant.

Figure 13 plots the Bayesian p-values for the 2PL model vs. those for the 3PL model. The performance of the 2PL model seems to be considerably worse than that of the 3PL model so far as item fit is concerned, resulting in 37 extreme p-values. The Rasch (1PL) model performs even worse, yielding 70 extreme p-values.

Item Fit Plots

Figure 14 shows the item fit plots for four items for the 3PL model fit to the data set. These plots are slightly different from the earlier item fit plots like Figure 10; the 90%
**Figure 13:** Posterior predictive p-values for the 2PL model vs. 3PL model fit to the admissions test data set.

The prediction interval for the proportions are shown using vertical dotted lines (and not using boxes, as before). The model seems to predict the responses for Item 16 satisfactorily. (The p-value for the χ²-type statistic for the item is 0.58.) There is considerable overall difference between the observed and replicated proportion corrects for Items 10, 33, and 65, with Item 33 appearing to be the worst. The p-values for the χ²-type statistic for these items are all less than 0.01. Of these two, Item 33 and 65 have lower proportion corrects than the model predicts for the high raw scores—that may have been caused by the items appearing towards the end of their respective sections and the presence of a speededness effect in the test, as observed by Sinharay (2003). Also, the plot for the Item 10 (an extremely difficult item) shows a strange behavior—the observed proportion correct for the item seems to decrease through the region from raw score of about 20 to raw score of 40 or so. The plot for this item seems to be similar to that for the “Bad Item 1” in Figure 8.
Figure 14: Item fit plots for four items for the 3PL model fit to the admissions test data set.

8. Conclusions

There exists no unanimously agreed upon statistical technique for assessing item fit for models in IRT. Measures suggested by Bock (1972), Yen (1981), Stone et al. (1994), and
Donoghue and Hombo (1999) are based on (pseudo-) counts that are actually not observed but estimated from the model, and hence they are not in the spirit of the traditional $\chi^2$ tests. Besides, these measures do not have well-established null distributions. The item fit measures suggested by Orlando and Thissen (2000) and Glas and Falcon (2003) are based on truly observed counts and have obvious intuitive appeal. However, the test by Glas and Falcon (2003) has undesirably high false alarm rates under certain situations and performs worse that the one by Orlando and Thissen (2000). On the other hand, the $\chi^2$ approximation of the statistics by Orlando and Thissen (2000) may be problematic, by arguments in Chernoff and Lehmann (1953); the departure of the test statistics from the assumed $\chi^2$ null distribution may be severe for short tests or small number of examinees. Simulation studies in Glas and Falcon (2003) and Orlando and Thissen (2000, 2003) show that the $\chi^2$ approximations of the $S - \chi^2$ and $S - G^2$ statistics may lead to occasionally high Type I error rates or false alarm rates. One way to establish the null distribution of the statistics is to perform extensive simulation studies, but that approach is problem specific, extremely computation intensive and it is extremely difficult, if not impossible, to account for all types of variation in this approach, making the approach impractical.

This paper shows how to use the PPMC method to assess item fit for the simple dichotomous IRT models. The technique suggested here uses the $\chi^2$-type and $G^2$-type test statistics of Orlando and Thissen (2000), but does not use an asymptotic $\chi^2$ approximation (nor does use extensive and inconclusive simulation studies) to determine the reference distribution. Instead, this paper uses the posterior predictive distribution (a natural one to use in Bayesian statistical analyses) of the test statistics as the reference distribution. The resulting p-values (PPP-value/Bayesian p-value) provide natural probability statements from a Bayesian viewpoint about the fit of the model to the items.

This paper also suggests item fit plots based on the posterior predictive checks. The observed proportion correct for individuals with different raw scores are compared graphically to the corresponding predictive distributions (posterior predictive distributions) suggested by the PPMC method.

A number of simulation studies demonstrate the usefulness of the item fit plot and the PPP-value. The PPP-values have Type I error rates and false alarm rates that never
exceed the nominal significance level and hence seem to be better than the existing item fit measures (because no test administrator would like to discard good items and the Bayesian tests ensure that). The measures seem to have considerable power for moderate to large sample sizes (power always very close to 1 for as low as 4,000 examinees for the situations examined). The simulations indicate that the use of the Bayesian measures will potentially lead to considerable savings because they, while having adequate power, will flag good or borderline items less often than the existing item fit measures.

It is important to apply any suggested technique to real data sets because they often reveal weaknesses of a technique that simulated data sets do not. Therefore, the suggested techniques are applied to three real data examples. There are not many published real data examples regarding item fit measures; therefore, these examples serve an important purpose. The results obtained from the Bayesian techniques are close to those obtained from the techniques of Orlando and Thissen (2000) for moderate or large data sets (Examples 2 and 3), but markedly different for a small number set (Example 1), where the $\chi^2$ approximation suggested by Orlando and Thissen (2000) is far from the truth. The examples show that the Bayesian diagnostics seem to be satisfactory for real data sets as well; they have considerable power for small to moderate real data sets.

The Bayesian approach seems to eliminate the limitations of the existing item fit techniques in that they not only have low Type I error rate, but seem to have acceptable power. The PPP-values are especially advantageous over their closest competitors, those by Orlando and Thissen (2000), for a small number of items and/or small number of examinees, when the $\chi^2$ approximation in Orlando and Thissen (2000) will lead to a test that suggests misfit too often. For large number of examinees and items, the null distribution of the $S - \chi^2$ and $S - G^2$ statistics will be well-approximated by $\chi^2$ distribution. However, it is extremely difficult to determine beforehand if the approximation will have acceptable Type I error rate for the specific application that a researcher is interested in. On the other hand, the Bayesian p-values, known to be conservative, will have acceptable Type I error rate irrespective of the number of the items or examinees and provide a reasonable way for item fit diagnostics. They are conservative, but that means that by rejecting borderline items less often, they may save money for the test administrators. The suggested techniques are
based on the MCMC algorithms, but that should not be a deterrent, given the recent surge in the use of the algorithm in psychometrics, the easy availability of software applying the MCMC algorithm in the field, and the advent of faster computational facilities.

Assessing item fit then is an example where a Bayesian approach provides a satisfactory solution to a problem where the frequentist approach is often found inadequate.
References


