A Posterior Predictive Model Checking Method Assuming Posterior Normality for Item Response Theory

Applied Psychological Measurement
2019, Vol. 43(2) 125–142
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621618779985
journals.sagepub.com/home/apm

Megan Kuhfeld¹

Abstract

This study investigated the violation of local independence assumptions within unidimensional item response theory (IRT) models. Bayesian posterior predictive model checking (PPMC) methods are increasingly being used to investigate multidimensionality in IRT models. The current work proposes a PPMC method for evaluating local dependence in IRT models that are estimated using full-information maximum likelihood. The proposed approach, which was termed as “PPMC assuming posterior normality” (PPMC-N), provides a straightforward method to account for parameter uncertainty in model fit assessment. A simulation study demonstrated the comparability of the PPMC-N and the Bayesian PPMC approach in the detection of local dependence in dichotomous IRT models.

Keywords

item response theory, posterior predictive model checking, model fit assessment

Introduction

Traditional item response theory (IRT) models specify a single continuous latent variable θ that explains the probability of endorsing a set of dichotomous or polytomous item responses, with the goal of representing individual differences on a psychological, behavioral, or educational construct. Accurate estimation of item parameters and the inferences made based on the estimates of θ depend on the degree to which the unidimensionality assumption holds in the observed item response data. The key implication of unidimensionality within IRT is the conditional (or local) independence assumption. Local independence means that for a given value of the latent variable θ , the joint probability of correct responses to an item pair is the product of the probabilities of correct responses to the two items.

However, there are many situations where this assumption is unlikely to hold. For example, items within a reading assessment that follow the same passage are likely to exhibit local dependence (LD). Clusters of items with overly similar meaning or phrasing may also violate the assumption of unidimensionality. In addition, many psychological assessments not only are designed to measure an overall dimension, but also are intended to measure subdomains, which would result in

¹The University of Texas at Austin, USA

Corresponding Author:

Megan Kuhfeld, 121 NW Everett Street, Portland, OR 97209.
Email: megan.kuhfeld@gmail.com

LD if the subdimensions were not explicitly modeled. Ignoring LD can result in biased item parameter estimates (Chen & Thissen, 1997; Tuerlinckx & De Boeck, 2001). As a result, LD can lead to poor estimation of the individual latent construct (Zenisky, Hambleton, & Sireci, 2002) and over estimation of the IRT information and test reliability (Sireci, Thissen, & Wainer, 1991).

Previous researchers have distinguished between types of LD present among items, which have differential effects on parameter estimation (Chen & Thissen, 1997). The two types of LD are *underlying* local dependence (ULD), which results from unmodeled latent traits, and *surface* local dependence (SLD), which is due to highly similar content or placement. Surface LD is defined by the probability π_{LD} that the second item in the LD pair will have a response identical to the first item. That is to say, for a set portion of respondents given by π_{LD} , the second response is determined completely by the first item within the item pair ($y_2 = 1$ if $y_1 = 1$, $y_2 = 0$ otherwise). SLD is generally seen as nuisance, whereas ULD can be seen as nuisance dimensions or as evidence of the need for latent dimensions to be added to the model.

There are two parallel lines of research for detecting LD that are used in the context of the two different IRT parameter estimation methods: frequentist (maximum likelihood) and Bayesian estimation. A full comparison of the theory, rationale, and assumptions of maximum likelihood and Bayesian estimation methods is beyond the scope of this study (see Cai & Thissen, 2015, and Fox, 2010, for overviews). Instead, comparing and contrasting the model fit procedures specifically related to the examination of LD within each estimation framework was focused.

Numerous indices and statistics for detecting LD have been proposed for IRT models estimated using maximum likelihood. Houts and Edwards (2013) provide an overview of the commonly used IRT-based tests of LD, including Yen's (1984) Q_3 statistic, Chen and Thissen's (1997) use of the Pearson's χ^2 and the likelihood ratio statistic G^2 , and Ip's (2001) suggestion to use the Mantel-Haenszel (MH) test with multiple testing corrections. These statistics are estimated based on a process that involves fitting a unidimensional IRT model to the data, and then comparing the observed responses with the model-expected responses.

Within Bayesian estimation, there is a promising set of model assessment techniques involving posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Meng, 1994). This model checking method requires the researcher to simulate draws from the full posterior distribution of the model parameters, which is straightforward when one uses Bayesian sampling-based methods (e.g., Markov chain Monte Carlo (MCMC) estimation; Gilks, Richardson, & Spiegelhalter, 1996). Sinharay, Johnson, and Stern (2006), Levy, Mislevy, and Sinharay (2009), Levy and Svetina (2011), among others, have demonstrated the PPMC method to assess LD in IRT models using various discrepancy measures, including the Pearson's χ^2 , MH statistic, and the likelihood ratio G^2 .

Although the same general set of discrepancy measures are used in the maximum likelihood and Bayesian examination of LD, the procedures and underlying assumptions are quite different. Frequentist approaches for assessing model fit evaluate the discrepancy between the observed data and the hypothesized model when the unknown model parameters are replaced by the best-fitting point estimates. There are two central problems with this approach. First, the use of point estimates for the unknown parameters in model fit assessment understates the uncertainty in the sampling distributions of discrepancy measures (Meng, 1994). Second, the null sampling distributions for most discrepancy statistics in likelihood model fit approaches are justified only asymptotically, which may not be tenable in real data situations.

By contrast, the PPMC method is a simulation-based model checking method that does not require asymptotic arguments. When the plausible values of the data can be drawn from the posterior predictive distribution, empirically constructing reference distributions of test quantities does not present a problem. Furthermore, PPMC integrates parameter uncertainty into model fit assessment through the use of the posterior predictive distribution.

The PPMC method has multiple advantages over frequentist approaches to check for the local independence. However, maximum marginal likelihood methods for parameter estimation and model fit testing currently dominate the field of IRT models, and the adoption of fully Bayesian approaches is likely to introduce many new complexities for applied users of IRT. Without additional training, both the specification of prior distributions and convergence monitoring of MCMC may present large barriers to an applied researcher. The goal of this study is to develop and evaluate a method that allows researchers to reap benefits of the (inherently Bayesian) PPMC method while using likelihood estimation.

The current study expands upon an alternative PPMC method that was developed in the context of structural equation modeling (Lee, Cai, & Kuhfeld, 2016). This method, which is termed the Poor Person's PPMC (PP-PPMC) method, employs only byproducts of maximum likelihood estimation (e.g., the maximum likelihood parameter estimates and the associated asymptotic covariance matrix) in the estimation of the replicated data. This method is based on a well-known result in Bayesian literature that, asymptotically, the likelihood tends to dominate the posterior shape. This study expands the PP-PPMC method into the context of IRT. The focus of the current study is to demonstrate the extent to which the use of the multivariate normal approximation of the parameter posterior in PPMC methods leads to similar results as the Bayesian PPMC method. The remainder of this article is organized as follows. First, the original PPMC method is introduced to provide the necessary background for discussing the proposed PPMC assuming posterior normality (PPMC-N) method. The remaining sections describe the results of a Monte Carlo study to test for LD using the various PPMC approaches. A real data application is provided, and the article is concluded with discussions about practical implications of the proposed method and future research.

PPMC Method

The PPMC method is a widely used Bayesian model checking method because it not only has a strong theoretical basis but is also easy to implement when conducting Bayesian estimation (Sinharay et al., 2006). The method primarily consists of comparing the observed data with replicated data (those predicted by the model) using a number of discrepancy measures. Systematic discrepancies between the observed data set and the replicated data sets indicate that the hypothesized model is failing to explain specific aspects of the observed data.

The observed data are denoted as $\mathbf{y}^{obs}$, whereas the replicated data are denoted as $\mathbf{y}^{rep}$. In the Bayesian framework, the likelihood $p(\mathbf{y}^{obs}|\boldsymbol{\gamma})$ is defined for a model applied to the data $\mathbf{y}^{obs}$, where $\boldsymbol{\gamma}$ denotes the set of parameters in the model. Let $p(\boldsymbol{\gamma})$ denote the prior distribution on the parameters. The posterior distribution of the parameters given hypothesized model H is

$$p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H) \propto p(\mathbf{y}^{obs}|\boldsymbol{\gamma}, H)p(\boldsymbol{\gamma}|H). \quad (1)$$

The hypothetical replicated data can be simulated from the posterior predictive distribution, which is the conditional distribution of the replicated data $\mathbf{y}^{rep}$ given the observed data $\mathbf{y}^{obs}$ and the model being tested, H . The general form of the posterior predictive distribution is

$$p(\mathbf{y}^{rep}|\mathbf{y}^{obs}, H) = \int_{\boldsymbol{\gamma}} p(\mathbf{y}^{rep}|\boldsymbol{\gamma}, H)p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)d\boldsymbol{\gamma}. \quad (2)$$

In Equation 2, there are two components: (a) $p(\mathbf{y}^{rep}|\boldsymbol{\gamma}, H)$: the sampling distribution of the replicated data given the hypothesized model and parameters, and (b) $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$: the posterior distribution of the model parameters for the observed data.

Implementation of the PPMC method involves multidimensional integration to obtain the posterior predictive distribution (see Equation 2), which in most circumstances is analytically intractable. Therefore, Monte Carlo simulation methods are employed in practice. First, a large number (L) of sets of plausible parameters $\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^L$ are drawn from the posterior distribution of the parameters. Second, For each plausible parameter vector $\boldsymbol{\gamma}^\ell$, one hypothetical replicate data set $\mathbf{y}^{rep, \ell}$ is drawn from the sampling distribution $p(\mathbf{y}^{rep, \ell} | \boldsymbol{\gamma}^\ell, H)$. Third, these L pairs of draws ($\boldsymbol{\gamma}^\ell$ and $\mathbf{y}^{rep, \ell}$) are used to estimate the degree to which the observed data $\mathbf{y}^{obs}$ and the replicated data $\mathbf{y}^{rep}$ are discrepant using appropriate discrepancy measures (Gelman et al., 1996).

Discrepancy measures, also known as test quantities $T(\mathbf{y}, \boldsymbol{\gamma})$, are defined to examine an aspect of discrepancy that is of interest. The arguments of the discrepancy measure are a data set (observed or replicated) and the model parameters. The test quantity $T(\mathbf{y}^{obs}, \boldsymbol{\gamma})$, where the data are fixed to the observed values but $\boldsymbol{\gamma}$ is drawn from the parameter posterior $p(\boldsymbol{\gamma} | \mathbf{y}^{obs}, H)$, is referred to as the *realized* test quantity. The *predictive* test quantity, $T(\mathbf{y}^{rep}, \boldsymbol{\gamma})$, is based on the joint posterior distribution of the replicated data $\mathbf{y}^{rep}$ and $\boldsymbol{\gamma}$. The *realized* and *predictive* test quantities are compared, with any significant difference between the test quantities indicating a failure of the model.

The PPMC method is primarily used to do graphical checks to compare the realized and predictive discrepancy measures. In addition, a Bayesian counterpart of the classical p value, known as the posterior predictive p value (PPP), can be found by calculating the tail area probability of the distribution in Equation 3:

$$p[T(\mathbf{y}^{rep}, \boldsymbol{\gamma}) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}) | \mathbf{y}^{obs}, H] = \int_{T(\mathbf{y}^{rep}, \boldsymbol{\gamma}) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma})} p(\mathbf{y}^{rep} | \boldsymbol{\gamma}, H) p(\boldsymbol{\gamma} | \mathbf{y}^{obs}, H) d\mathbf{y}^{rep} d\boldsymbol{\gamma}. \quad (3)$$

That is, the Bayesian PPP value is defined as the probability that the predictive discrepancy measure is more extreme than the realized discrepancy measure. When the fitted model is correct, the PPP values are not necessarily uniformly distributed, and previous researchers have found the PPP values for a correct model tend to be closer to .5 (Sinharay & Stern, 2003). Extremely small probabilities (less than .05 or greater than .95) provide a strong suggestion that the model is not capturing the necessary features of the data.

Alternatives to Posterior Distribution in Predictive Model Checking

Various alternatives have been proposed to the parameter posterior distribution $p(\boldsymbol{\gamma} | \mathbf{y}^{obs}, H)$ in the construction of the posterior predictive distribution (Gelfand, 1996). Box (1980) suggested the use of the *prior* distribution (prior predictive model checking), whereas Bayarri and Berger (2000) proposed the use of the *conditional posterior* distribution (partial or conditional PPMC method). Levy (2011) has previously discussed the similarities between posterior predictive, prior predictive, and partial predictive model checking. Sinharay (2016) described a generalized resampling-based approach for person-fit assessment (PFA), which encompasses various resampling methods of the ability distribution.

Furthermore, Lee et al. (2016) described a general expression under which all of the previously suggested predictive distributions can be combined:

$$p(\mathbf{y}^{rep} | \mathcal{A}, H) = \int_{\boldsymbol{\gamma}} p(\mathbf{y}^{rep} | \boldsymbol{\gamma}, H) p(\boldsymbol{\gamma} | \mathcal{A}, H) d\boldsymbol{\gamma}. \quad (4)$$

The general expression becomes the prior predictive distribution if $\mathcal{A}$ is an empty set, whereas the general expression becomes the posterior predictive distribution (Equation 2) if $\mathcal{A}$ is the observed data set $\mathbf{y}^{obs}$.

PPMC With a Normality Assumption

Building off this general expression, Lee et al. (2016) suggested the use of a multivariate normal distribution with its mean vector equal to the maximum likelihood estimate $\hat{\boldsymbol{\gamma}}$ and dispersion matrix equal to the asymptotic covariance matrix. This choice was motivated by the fact that the likelihood tends to dominate the posterior shape asymptotically (Gelman, Carlin, Stern, & Rubin, 2003). As a result, the posterior distribution, to a first approximation, converges to a multivariate normal with its mean centered around $\hat{\boldsymbol{\gamma}}$.

In place of $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$ in Equation 2, the PPMC with normality assumption uses the multivariate normal distribution with its mean vector equal to $\hat{\boldsymbol{\gamma}}$ and dispersion matrix equal to the asymptotic covariance matrix of the maximum likelihood estimate $\hat{\mathbf{V}}$

$$\phi_d(\boldsymbol{\gamma}) = |2\pi\hat{\mathbf{V}}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})' \hat{\mathbf{V}}^{-1}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\right\}, \quad (5)$$

where $\phi_d(\boldsymbol{\gamma})$ is the multivariate normal distribution with d dimensions. That is to say, this proposed PPMC approach relies on posterior normality assumption to construct an approximate posterior predictive distribution using only byproducts of maximum likelihood estimation. For brevity, this proposed approach is subsequently referred to PPMC-N, where the ‘‘N’’ denotes the posterior normality assumption.

The only modification from the PPMC procedure described previously is that the parameter posterior distribution is replaced with the multivariate normal approximation $\phi_d(\boldsymbol{\gamma})$. The subsequent steps of drawing replicate data sets and estimating realized and predictive test quantities remain the same. This approach does not require conducting Bayesian data analysis, but still incorporates parameter uncertainty in model assessment through the draws of $\boldsymbol{\gamma}^\ell$. Using the predictive and realized test quantities estimated with the PPMC-N approach, the proportion of draws for which the predictive test quantity exceeds its corresponding realized test quantity can be calculated. These PPMC-N approximated Bayesian p values are called PPMC-N predictive p values:

$$\text{PPMC-N predictive pvalue} \approx \frac{1}{L} \sum_{\ell=1}^L \mathbf{1}\{T(\mathbf{y}^{rep, \ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\}, \quad (6)$$

where $\mathbf{1}\{T(\mathbf{y}^{rep, \ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\}$ is an indicator function that takes a value of 1 only if $\{T(\mathbf{y}^{rep, \ell}, \boldsymbol{\gamma}^\ell) \geq T(\mathbf{y}^{obs}, \boldsymbol{\gamma}^\ell)\}$ is true.

PPMC With a Normality Assumption and Resampling

A concern with the PPMC-N approach is that it is dependent upon a multivariate normality approximation for the posterior of the item parameters $p(\boldsymbol{\gamma}|\mathbf{y}^{obs}, H)$. It is not clear how realistic this assumption is with real data. The sampling importance resampling (SIR) approach (Rubin, 1987) is proposed as a potential method to address this concern. Importance sampling is a Monte Carlo method where a target distribution is approximated by a weighted average of random draws from another distribution. In our case, the target distribution from which we want to

sample is the parameter posterior distribution $p(\boldsymbol{\gamma}^\ell | \mathbf{y}^{obs})$, and $\phi_d(\boldsymbol{\gamma}^\ell)$ is a proposal distribution from which samples can be drawn.

For the PPMC-N method, the SIR approach breaks the posterior predictive simulations into several steps. First, L samples are drawn from multivariate normal distribution centered at maximum likelihood estimates $\hat{\boldsymbol{\gamma}}$ and dispersion $\hat{\mathbf{V}}$ equal to inverse Fisher information matrix. The parameter posterior function in Equation 1 is evaluated based on the current set of $\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^L$ draws. Then, importance ratios are calculated by dividing the parameter posterior distribution by the multivariate normal density at each of the L draws.

$$\omega_\ell = \frac{p(\boldsymbol{\gamma}^\ell | \mathbf{y}^{obs})}{\phi_d(\boldsymbol{\gamma}^\ell)} \quad (7)$$

The importance ratios are normalized ($q_\ell = \omega_\ell / \sum_{m=1}^L \omega_m$), and then a new set of parameters $\{\boldsymbol{\gamma}^{1*}, \dots, \boldsymbol{\gamma}^{L*}\}$ is drawn from the discrete distribution over $\{\boldsymbol{\gamma}^1, \dots, \boldsymbol{\gamma}^L\}$ placing mass q_ℓ on $\boldsymbol{\gamma}^\ell$. For each resampled parameter draw, a predictive draw of replicate data is produced. Finally, the predictive and realized test quantities are estimated and compared. This combined approach of PPMC-N with SIR re-sampling is termed PPMC-N-SIR.

Method

The purpose of this study is to evaluate different predictive model checking approaches using a simulation study, specifically focusing on detecting violations of unidimensionality within the two-parameter logistic (2PL) IRT model. The independent variables considered in this simulation study were (a) sample size ($N = 250, 500, 1,000$), (b) data-generating models (no LD, SLD, and ULD), and (c) within the LD conditions, SLD values were $\pi_{LD} = 0.5$ and 0.8 , and ULD values were slope ratios of 0.5 and 1.5 . In all of the conditions, there are 20 items. One hundred replications were conducted within each condition.

Discrepancy Measures

The current work focuses on bivariate discrepancy measures to test for LD within pairs of items. The discrepancy measures should be chosen to reflect important features of the model and reveals useful information about when model assumptions are violated. This study examines eight bivariate discrepancy measures that have been previously utilized by Levy (2006) to detect multidimensionality in IRT models: χ^2 , G^2 , Q_3 , sample covariance, model-based covariance (MBC), residual covariance, natural log of the odds ratio, and a standardized log odds ratio residual. The equations for these discrepancy measures are given in the appendix.

Data Generation

True latent ability (θ) were generated from a standard normal distribution in the unidimensional and SLD conditions, and a multivariate normal distribution in third data-generating condition. Item discrimination parameters ranged from values of 1 to 2, and item intercepts were set to values of $-1, -.5, .5, 1$, or 1.5 .

In the null condition, data were generated following a unidimensional model. Programs were written in R (version 3.1) to simulate data in the null, SLD, and ULD conditions. Surface LD was generated following the deterministic process outlined in the Introduction section. The probability that the second item within a pair will have the same response, π_{LD} , is manipulated

to vary the degree of LD. The two values of π_{LD} considered are 0.5 and 0.8. There are four LD item pairs in the surface LD condition, and the remaining 16 items are locally independent.

For the underlying LD condition, data were generated following a bifactor model (Holzinger & Swineford, 1939). The bifactor model contains a general factor that explains the overall variance among the items and specific factors that account for residual variance. The specific factors are uncorrelated with each other and the general factor. Each specific factor contains five items, with a total of four specific factors. For items within each specific factor, the degree of LD was defined by the ratio of the general slope to the specific factor slopes. The specific factor slopes were either set to be half of the general slopes, or set to be 1.5 times the general factor slopes. The variances of all latent variables in the bifactor model were set to 1.

Parameter Estimation

For each replication, item parameters were estimated using a unidimensional 2PL IRT model. For the two PPMC approaches that rely on normality approximations, parameters were estimated using Bock–Aitkin expectation-maximization (EM) algorithm implemented in flexMIRT[®] (Cai, 2015).

In addition, a fully Bayesian IRT calibration was conducted using MCMC estimation with the freeware WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). The 2PL model was estimated using standard normal prior distribution for θ and b_k (see Appendix for more details), and a truncated normal prior distribution for a_k . A total of 2,000 iterative simulations were done per chain (three total chains), the first 400 iterations were discarded, and the iterations were thinned by a value of 2. After discarding burn-in and thinning the chain, the resulting iterations were pooled to produce a total set of iterations for use in PPMC.

PPMC Estimation

For each replication, PPMC, PPMC-N, and PPMC-N-SIR predictive p values were calculated. An R program was written to generate simulated data sets and calculate the discrepancy measures. The PPMC approach utilized $L = 100$ draws from the posterior distribution of the item parameters and the posterior distribution of the latent variables q . Based on these estimates, replicate data were drawn and the predictive p values were estimated for each of the discrepancy measures. Various numbers of draws ($L = 100, 500, 1,000$) were compared, with no discernible difference in results. Therefore, $L = 100$ was chosen to reduce the computational time.

For the PPMC-N and PPMC-N-SIR methods, an R program was written to read in the flexMIRT item parameter estimates $\hat{\gamma}$ and the asymptotic covariance matrix and to simulate random normal draws. The PPMC-N method utilized $L = 100$ draws from the multivariate normal approximation to the parameter posterior distribution. For the PPMC-N-SIR approach, an initial sample of $L = 1,000$ parameter sets were drawn, and then using the SIR reweighting, a final set of $L = 100$ draws were used to calculate the PPMC-N-SIR predictive p values.

Results

Null Conditions

A central question of this study is whether the multivariate normal distribution is a reasonable approximation for the posterior parameter distributions. Figure 1 displays the estimated marginal posterior density, the multivariate normal posterior approximation, and the multivariate normal posterior approximation with resampling for the 20 item slope parameters estimated in a

Figure 1. Comparison of three predictive model checking methods in the approximation of the item parameter marginal posterior distributions within a single replication, null condition ($N = 250$).

Note. * = true parameter; PPMC = posterior predictive model checking; PPMC-N = posterior predictive model checking assuming normality; PPMC-N-SIR = posterior predictive model checking assuming normality using SIR.

single replication under the null condition ($N = 250$). The true generating item parameters are shown as an asterisk on the x -axis. The two posterior approximations closely overlap with the marginal parameter posterior distributions estimated using MCMC, and all three approaches are generally doing an excellent job of recovering the true parameters. In the cases where the posterior and normal approximation are discrepant, such as the first and fifth slope parameter, the distributions for the SIR approach fall in between the marginal multivariate normal approximation and the posterior distributions. Similar results are seen for the item intercepts. In the other sample size conditions, the distributions are narrower and peaked, demonstrating higher precision of the estimates. It is clear that even with small sample sizes, the multivariate normal approximation to the posterior is closely representing the posterior distribution in the null condition.

In addition, the performance of the eight discrepancy measures in the null conditions was examined. Given that the data were generated from a unidimensional model, all item pairs reflect the same single latent dimension, and the results were pooled across item pairs following

Figure 2. Comparison of three predictive model checking methods, null condition ($N = 250$).
 Note. OR = odds ratio; PPMC = posterior predictive model checking; PPMC-N = posterior predictive model checking assuming normality; PPMC-N-SIR = posterior predictive model checking assuming normality using SIR.

an exchangeability assumption (de Finetti, 1964). There are 20 items, and therefore there are $(20 \times (20 - 1)) / 2 = 190$ item pairs. For the null condition, the predictive p values from the 190 item pairs for each bivariate discrepancy measure are pooled within a replication, and the results are then pooled across the 100 replications. Figure 2 contains eight panels, one for each discrepancy measure examined. In each panel, there are line plots for the distributions of the pooled p values showing the relative frequency of each p value (ranging from 0 to 1). The black line corresponds to the posterior predictive method, the blue dashed line corresponds to the PPMC-N method, and the red dotted line corresponds to the PPMC-N method with SIR.

Within the null condition, median values around 0.5 are expected, which is what is observed across all of the discrepancy measures in Figure 2. The three predictive approaches produce mostly overlapping distributions, with the exception of χ^2 and G^2 . For these three discrepancy measures, the PPMC-N and PPMC-N-SIR approaches are shifted slightly closer to 1. As observed by others (Meng, 1994), the distributions of the predictive p values for all the discrepancy measures are less dispersed than a uniform distribution. This is seen particularly for χ^2 and G^2 , where the predictive p values are heavily concentrated between .3 and .7. The discrepancy measures displaying the most uniform distribution were Q_3 and the MBC matrix.

To further describe the distribution of the predictive p values under the null condition, the proportion of extreme predictive p values was examined. Extreme predictive p values are defined as those below .05 or above .95, which can be considered as Type I error rates using .05 and .95 as critical values (i.e., in a two-tailed test with $\alpha = .10$). The empirical Type I error rates for the covariance and residual covariance discrepancy measures are close to the nominal

rate of .10. However, empirical Type I error rates for the other discrepancy measures are below .10, indicating that use of predictive p values in hypothesis testing results in a conservative test. The results under the null condition are highly similar for the other sample sizes.

LD Conditions

The next set of results focuses on the predictive model checking approaches where LD is present in a subset of items. In the multidimensional conditions, there are item pairs that contain surface or underlying LD, and item pairs that are not misfit in each condition. The results for item pairs that are not misfit will not be presented. The median predictive p values for the nonmisfit item pairs are around .5 and there are not systematic patterns across conditions.

As with the unidimensional data, predictive p values obtained from each discrepancy measure on multiple misfit item pairs are pooled following exchangeability assumptions, as well as across the 100 replications within conditions. Table 1 presents the median predictive p values for the misfit items for all of the eight examined discrepancy measures across the four LD conditions. The first two columns list the type of LD misspecification and the predictive model checking method. The performance of the various discrepancy measures within a LD condition is consistent across sample size and model checking approach. Results within a condition also appear fairly similar across the discrepancy measures, indicating little is gained by using all eight different discrepancy measures. The notable exception is that in the mild conditions, the median values of the χ^2 and G^2 discrepancy measures are quite close to one another (and are frequently equal) and generally the closest to .5. In the mild ULD condition, Q_3 and the MBC have the highest median p values. All of the discrepancy measures perform almost equally in the strong misfit conditions.

Table 2 presents the proportion of replications with extreme p values for the misfit items ($p < .05$ or $> .95$) for all of the eight studied discrepancy measures. Again, the performances of the different discrepancy measures within a LD condition are consistent across sample size and model checking approach. Again, the Q_3 and the MBC discrepancy measures demonstrate the highest proportion extreme p values within the misfit items, indicating they are most sensitive to mild levels of LD misfit. However, all of the discrepancy measures behave similarly in the strong LD conditions.

Application to Real Data

The data used in this example were collected as a part of the Child and Young Adult supplement to the National Longitudinal Study of Youth 1979. The Child and Young Adult supplement constitutes a sample of children of women who had participated in the National Longitudinal Study of Youth longitudinal study, a nationally representative study designed by the United States Department of Labor to study variations in labor market behavior and experiences. Beginning in 1986, interviewers administered an extensive set of assessment instruments to the children of all female respondents. These assessments included information about cognitive, socioemotional, and psychological aspects of the child's development as well as about the quality of the home environment. In 2-year increments from 1986 to 2012, the Peabody Individual Achievement Test (PIAT; Dunn & Markwardt, 1970) was administered to children in the study who were between the ages of 5 and 15 years at each wave. The PIAT Math test consists of 84 multiple-choice items designed to measure mathematics concepts taught in mainstream classrooms. The problems were designed so that children are required to apply math concepts to questions rather than conduct increasingly complicated computations. The test starts with basic skills, such as number recognition and counting, and increases in difficulty to problems involving division, multiplication, and fractions.

Table 1. Median p Values for the Eight Discrepancy Measures for the Misfit Items Within the Four LD Conditions.

LD condition	PPMC method	Sample size	χ^2	G^2	Q_3	Cov	MBC	Residual cov	LN(OR)	STDLN(OR)
SLD—mild	PPMC-N	250	0.54	0.54	0.59	0.55	0.58	0.55	0.55	0.55
	PPMC-N-SIR	250	0.51	0.51	0.59	0.60	0.60	0.60	0.60	0.59
	PPMC	250	0.48	0.48	0.56	0.46	0.57	0.46	0.46	0.46
	PPMC-N	500	0.52	0.52	0.57	0.56	0.57	0.56	0.56	0.56
	PPMC-N-SIR	500	0.52	0.53	0.56	0.55	0.56	0.55	0.55	0.55
	PPMC	500	0.47	0.48	0.55	0.48	0.55	0.48	0.48	0.48
	PPMC-N	1,000	0.49	0.49	0.44	0.42	0.43	0.42	0.42	0.41
	PPMC-N-SIR	1,000	0.50	0.51	0.43	0.41	0.43	0.41	0.41	0.41
	PPMC	1,000	0.47	0.47	0.40	0.32	0.40	0.32	0.32	0.32
SLD—strong	PPMC-N	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ULD—mild	PPMC-N	250	0.57	0.57	0.71	0.67	0.70	0.67	0.67	0.67
	PPMC-N-SIR	250	0.57	0.56	0.72	0.68	0.72	0.68	0.66	0.66
	PPMC	250	0.49	0.49	0.72	0.67	0.71	0.67	0.68	0.67
	PPMC-N	500	0.55	0.55	0.77	0.70	0.77	0.70	0.70	0.70
	PPMC-N-SIR	500	0.55	0.55	0.77	0.71	0.77	0.71	0.71	0.71
	PPMC	500	0.49	0.50	0.76	0.71	0.76	0.71	0.71	0.71
	PPMC-N	1,000	0.60	0.60	0.88	0.81	0.88	0.81	0.80	0.80
	PPMC-N-SIR	1,000	0.58	0.57	0.87	0.80	0.87	0.80	0.79	0.79
	PPMC	1,000	0.55	0.55	0.87	0.80	0.87	0.80	0.81	0.81
ULD—strong	PPMC-N	250	0.97	0.98	1.00	0.99	1.00	0.99	0.99	1.00
	PPMC-N-SIR	250	0.97	0.98	1.00	0.99	1.00	0.99	1.00	1.00
	PPMC	250	0.95	0.96	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. LD = local dependence; PPMC = posterior predictive model checking; cov = covariance; MBC = model-based covariance; residual cov = residual covariance; LN(OR) = natural log of the odds ratio; STDLN(OR) = standardized log odds ratio residual; SLD = surface local dependence; PPMC-N = posterior predictive model checking assuming normality; PPMC-N-SIR = posterior predictive model checking assuming normality using SIR; ULD = underlying local dependence.

For this application, the sample of children between the ages of 5 and 6 years who were administered the PIAT math test was used. These ages were selected because children enter the assessment at an age-appropriate item, and children within this age group all start at the first item. Only the first 15 items were selected for this analysis as the difficulty level of the later items was too high for this sample.

The final sample consists item responses of 3,385 children to the 15 items (scored as correct or incorrect). A 2PL model was estimated with the data. Table 3 shows the item parameters and the MBC discrepancy measure. Item sets that cover similar topic areas, such as the first four

Table 2. Proportion of Extreme p Values (i.e., p Value $< .05$ or $> .95$) for the Eight Discrepancy Measures Among the Misfit Items.

LD condition	PPMC method	Sample size	χ^2	G^2	Q_3	Cov	MBC	Residual cov	LN(OR)	STD LN(OR)
SLD—mild	PPMC-N	250	0.00	0.00	0.13	0.05	0.15	0.05	0.05	0.05
	PPMC-N-SIR	250	0.00	0.00	0.08	0.05	0.13	0.05	0.05	0.05
	PPMC	250	0.00	0.00	0.10	0.15	0.10	0.15	0.15	0.15
	PPMC-N	500	0.00	0.00	0.15	0.13	0.15	0.13	0.13	0.13
	PPMC-N-SIR	500	0.00	0.00	0.18	0.13	0.15	0.13	0.10	0.13
	PPMC	500	0.03	0.03	0.15	0.13	0.15	0.13	0.10	0.13
	PPMC-N	1,000	0.00	0.00	0.15	0.10	0.10	0.10	0.05	0.08
	PPMC-N-SIR	1,000	0.00	0.00	0.13	0.08	0.13	0.08	0.08	0.08
	PPMC	1,000	0.00	0.00	0.13	0.08	0.13	0.08	0.08	0.08
SLD—strong	PPMC-N	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	250	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	500	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ULD—mild	PPMC-N	250	0.01	0.01	0.15	0.05	0.14	0.05	0.06	0.07
	PPMC-N-SIR	250	0.00	0.00	0.16	0.04	0.15	0.04	0.05	0.05
	PPMC	250	0.00	0.00	0.14	0.05	0.14	0.05	0.04	0.04
	PPMC-N	500	0.02	0.01	0.16	0.08	0.17	0.08	0.07	0.07
	PPMC-N-SIR	500	0.02	0.02	0.20	0.07	0.19	0.07	0.08	0.07
	PPMC	500	0.01	0.01	0.18	0.07	0.18	0.07	0.08	0.08
	PPMC-N	1,000	0.03	0.03	0.29	0.13	0.28	0.13	0.13	0.14
	PPMC-N-SIR	1,000	0.02	0.03	0.28	0.14	0.28	0.14	0.14	0.14
	PPMC	1,000	0.02	0.02	0.30	0.13	0.30	0.13	0.15	0.15
ULD—strong	PPMC-N	250	0.63	0.66	0.94	0.85	0.94	0.85	0.85	0.86
	PPMC-N-SIR	250	0.60	0.62	0.93	0.84	0.93	0.84	0.84	0.84
	PPMC	250	0.47	0.50	0.94	0.87	0.94	0.87	0.87	0.87
	PPMC-N	500	0.94	0.94	1.00	0.99	1.00	0.99	0.99	0.99
	PPMC-N-SIR	500	0.95	0.95	1.00	0.98	1.00	0.98	0.99	0.99
	PPMC	500	0.90	0.91	1.00	0.98	1.00	0.98	0.98	0.98
	PPMC-N	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC-N-SIR	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	PPMC	1,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. LD = local dependence; PPMC = posterior predictive model checking; cov = covariance; MBC = model-based covariance; residual cov = residual covariance; LN(OR) = natural log of the odds ratio; STD LN(OR) = standardized log odds ratio residual; SLD = surface local dependence; PPMC-N = posterior predictive model checking assuming normality; PPMC-N-SIR = posterior predictive model checking assuming normality using SIR; ULD = underlying local dependence.

items that all involve identifying numbers, were flagged for LD using PPMC-N. Similar levels of LD are found for the set of items that involve counting objects (Items 7-9). Based on these results, multidimensional measurement models that account for the apparent math subdomains should be explored.

Conclusion

Motivated by the work on the Bayesian PPMC techniques for checking for unidimensionality in IRT models, this study proposes a normality approximation for the parameter posterior

Table 3. Item Parameter Estimates and PPMC-N Local Dependence Index for 15 PIAT Math Items.

N	Item	Percentage correct	c	a	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Find the number 4.	.99	6.08	1.6														
2	Find the number 2.	.98	5.11	1.57	1													
3	Find the number 3.	.98	5.7	1.76	1	1												
4	Find the number 5.	.97	4.43	1.51	1	1	1											
5	Find the biggest shape.	.97	4.06	1.1	1	1	1	1										
6	Find the shortest pencil.	.74	1.13	0.71	.9	.9	.3	.4	.5									
7	Tom has two cats and got one more. How many?	.78	1.58	1.25	.5	.3	.3	.6	.6	.8								
8	Mary had three dogs and gave away two. How many?	.85	2.55	1.72	.6	.4	.6	.6	.8	.4	1							
9	Look at the dogs and cats. Count how many.	.71	1.08	1.16	.5	.5	.7	.7	.7	.9	.9	1						
10	How many circles are shown?	.82	2.1	1.63	.4	.5	.6	.6	.9	.7	1	1						
11	Point to the curved lines.	.42	−0.37	0.72	.7	.4	.4	.2	.1	1	.8	.7	.1	.1				
12	Which number is twenty-eight?	.39	−0.59	1.04	.1	0	0	0	.2	.7	1	.4	.7	.3	.9			
13	Point to the circle that is cut in half.	.70	1.12	1.47	.3	.1	.5	.3	.7	.9	.6	1	1	.9	.5	.1		
14	Janet had five pennies and spent three. How many?	.38	−0.65	1.22	.1	.4	.1	.5	.3	.5	.7	.7	.8	.9	.1	.2	.2	
15	Which is the youngest kid's birthday cake?	.40	−0.64	1.57	.2	.5	.5	.1	.4	1	.3	.6	.6	.6	.6	.3	.4	.2

Note. PPMC-N = posterior predictive model checking-normality; PIAT = Peabody Individual Achievement Test.

distribution that allows a predictive model checking approach to be implemented under maximum likelihood estimation. A PPMC method assuming posterior normality (PPMC-N) for IRT models is proposed to take advantage of the flexibility of the previously studied PPMC approach (that is used in tandem with Bayesian estimation) when maximum likelihood estimation methods are used. The PPMC-N method integrates parameter uncertainty into model fit assessment through the use of the random draws from the multivariate normal parameter distribution, and does not rely on asymptotically defined distributions for the discrepancy statistics.

Using a set of discrepancy measures that have been previously identified for the detection of multidimensionality in IRT, a simulation study compared the performance of (a) the fully Bayesian PPMC method; (b) the PPMC-N method, in which the posterior predictive distribution is approximated using a multivariate normal distribution that is centered around the maximum likelihood estimates of the parameters; and (c) the PPMC-N-SIR method, which relies on an additional SIR step to decrease the reliance on the multivariate normal approximation of the posterior. Key findings are that the normal approximation to the marginal parameter posteriors closely matched the MCMC-estimated posterior distribution and that results from the three PPMC methods were similar across a range of LD conditions and sample sizes. This finding is parallel to the findings of Sinharay (2016) in the context of PFA, which demonstrated that the Monte Carlo resampling method and bootstrap method and the parametric bootstrap method resulted in similar conclusions.

All of the bivariate discrepancy measures examined performed similarly in the strong LD conditions, showing excellent detection of the LD items. The most effective measures in the mild LD conditions were Q_3 and the MBC, which performed almost identically. In addition, in the null condition, these discrepancy measures displayed distributions of predictive p values that were closest to uniform. These discrepancy measures are recommended for detecting LD using PPMC-N. These results parallel the findings of prior researchers, including Levy et al. (2009). Further research will be needed to examine whether the PPMC-N approach is sensitive to negative LD between items that reflect different underlying dimensions (Habing & Roussos, 2003; Levy, 2006).

The results from this simulation study are generalizable only to the extent that the design variables are similar to real-world data conditions. Sample sizes were chosen to be reflective of small to large samples that may be used in IRT calibration, but the simulated data had no missing responses, which is likely in real-world applications. Further work is necessary to examine the PPMC-N approach with shorter item banks and other item response models (e.g., models for polytomous item responses or models that contain a guessing parameter). It is possible that the normal approximation used in this approach will not work as well for lower-asymptote parameters.

The goal of this study was to provide additional ways for applied users of IRT, who typically rely on maximum likelihood methods of estimation available in widely used software, to detect violations of model assumptions. However, subjective judgments on the part of the user are still required when deciding whether the discrepancies between the data and hypothesized are important enough to address. When misfit is detected, possible actions include (a) retaining the hypothesized model to make inferences but acknowledging the limitations and source of misfit, (b) discarding locally dependent items and retesting the model with the revised data, or (c) using a more general model that accounts for the extra dimensionality (such as a bifactor or testlet model). These decisions remain as an important piece of model appraisal, which can be best made with strong information regarding the specific aspects of the model that are discrepant with the data. The PPMC-N method is a flexible alternative to both traditional ML and Bayesian model checking methods, and retains many of the advantages of the fully Bayesian approach while remaining a viable options for those who choose the use maximum likelihood estimation methods.

Appendix

Discrepancy Measures

Discrepancy measures that examine the pairwise association of items can be used to detect local item dependence. Many of the bivariate discrepancy measures involve observed and expected frequencies of item responses for two items (k and k') at a time (e.g., two-way tables for the frequencies of y_k and $y_{k'}$). The observed bivariate table for a dichotomous item is

		y_k	
		1	0
$y_{k'}$	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

where n_{11} is the number of examinees with a correct response for both y_k and $y_{k'}$. The expected bivariate table is

		y_k	
		1	0
$y_{k'}$	1	$E(n_{11})$	$E(n_{10})$
	0	$E(n_{01})$	$E(n_{00})$

where $E(n_{01})$ is the expected frequency of a correct response to y_k and an incorrect response to $y_{k'}$ predicted by the item response theory (IRT) model. Due to assumptions of local independence, $E(n_{01})$ can be estimated as follows:

$$E(n_{01}) = N \int P_k(\theta)[1 - P_{k'}(\theta)]p(\theta)d\theta, \quad (1)$$

where $P_k(\theta)$ is the traceline for item k , $P_{k'}(\theta)$ is the traceline for item k' , $p(\theta)$ is the population distribution, and N is the total number of examinees. This integral can be approximated numerically using a set of quadrature points.

The χ^2 and G^2 discrepancy measures for item pairs (Chen & Thissen, 1997) are given, respectively, by

$$\begin{aligned} \chi^2_{kk'} &= \sum_{c=0}^1 \sum_{c'=0}^1 \frac{(n_{cc'} - E(n_{cc'}))^2}{E(n_{cc'})}, \\ G^2_{kk'} &= -2 \sum_{c=0}^1 \sum_{c'=0}^1 n_{cc'} \ln \frac{E(n_{cc'})}{n_{cc'}}. \end{aligned} \quad (2)$$

In addition, Yen's (1984, 1993) Q_3 statistic measures the correlation between a pair of items after accounting for the latent traits. The deviation between the observed and expected response for the i th examinee is calculated as

$$d_{ki} = y_{ki} - E_{y_{ki}}, \quad (3)$$

where y_{ki} is the observed response to item k and $E_{y_{ki}}$ is that the examinee's expected response to item k . For dichotomous items, the expected response can be calculated as $E_{y_{ki}} = P(y_{ki} = 1 | \theta_i, a_k, c_k)$. The LD index Q_3 is computed as the Pearson product-moment correlation between the deviation scores of items k and k' .

The model-based covariance (MBC) discrepancy measure, which also examines deviation between observed and expected responses, was proposed by Reckase (1997).

$$\text{MBC}_{kk'} = \frac{\sum_{i=1}^N (y_{ki} - E_{y_{ki}})(y_{k'i} - E_{y_{k'i}})}{N}. \quad (4)$$

For both Q_3 and Reckase's residual measure, a point estimate $\hat{\theta}$ is needed for each examinee. The *expected a posteriori* (EAP), or posterior mean, is a commonly used Bayesian estimator in IRT (Bock & Mislevy, 1982). The EAP estimator is calculated by taking the expectation over an individual's posterior distribution (e.g., product of the response pattern likelihood with a population distribution $p(\theta)$)

$$\hat{\theta}_i = \int_{\theta} \theta f(\theta | \mathbf{y}_i, \boldsymbol{\gamma}) d\theta = \frac{1}{f(\mathbf{y}_i | \boldsymbol{\gamma})} \int_{\theta} \theta L(\mathbf{y}_i | \theta, \boldsymbol{\gamma}) p(\theta) d\theta. \quad (5)$$

Typically, a standard normal distribution is used for $p(\theta)$ when estimating $\hat{\theta}$ scores. The sample covariance is another possible discrepancy measure and is given by

$$\text{COV}_{kk'} = \frac{[(n_{11})(n_{00}) - (n_{10})(n_{01})]}{N^2}. \quad (6)$$

Similarly, residual item correlations can be estimated, which compare the observed covariance with the expected covariance for an item pair

$$\text{ResidCOV}_{kk'} = \frac{[(n_{11})(n_{00}) - (n_{10})(n_{01})]}{N^2} - \frac{[E(n_{11})E(n_{00}) - E(n_{10})E(n_{01})]}{N^2}. \quad (7)$$

Finally, two odds ratio measures can be used as discrepancy measures. The odds ratio on the (natural) log scale is

$$\ln(\text{OR}_{kk'}) = \ln \left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})} \right]. \quad (8)$$

The standardized log ratio residual dependence formula given by Chen and Thissen (1997) is

$$\text{STD LN}(\text{OR}_{kk'}) = \frac{\ln \left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})} \right] - \ln \left[\frac{E(n_{11})E(n_{00})}{E(n_{10})E(n_{01})} \right]}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}}. \quad (9)$$

Details of the Bayesian Estimation Models

A fully Bayesian IRT calibration is conducted using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000). A unidimensional 2PL model is estimated in WinBUGS with the following prior distributions:

$$\theta_i \sim N(0.0, 1.0)$$

$$b_k \sim N(0.0, 1.0)$$

$$a_k \sim N + (1.0, 1.0)$$

where $N+(1.0, 1.0)$ denotes the normal distribution truncated to the positive real line. The threshold parameter b_k is estimated in the Bayesian model rather than the intercept c_k parameter, which can be easily converted back to the intercept to match flexMIRT's parameterization ($c_k = -a_k \times b_k$).

Convergence of these models was measured using the effective sample size (sample size adjusted for autocorrelation across simulations) and Gelman and Rubin (1992) convergence diagnostic (printed in R2WinBUGS as Rhat). In addition, plots of the Markov chain Monte Carlo (MCMC) sequences were examined for the parameter and θ estimates. The average Gelman and Rubin convergence diagnostic (Rhat) across parameters and conditions was 1.00.

Acknowledgments

The author thanks Li Cai, Qing Zhou, and Mark Hansen for their helpful advice on this manuscript.

Author's Note

Megan Kuhfeld is currently associated with NWEA.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models. *Journal of the American Statistical Association*, 95, 1127-1142.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431-444.
- Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society: Series A (General)*, 143, 383-430.
- Cai, L. (2015). flexMIRT® Version 3.51: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., & Thissen, D. (2015). Modern approaches to parameter estimation in item response theory. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 41-59). New York, NY: Routledge.
- Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22, 265-289.
- de Finetti, B. (1964). Foresight: Its logical laws, its subjective sources. In H. E. Smokler & H. E. Kyburg (Eds.), *Studies in subjective probability* (pp. 93-158). New York, NY: Wiley.
- Dunn, L. M., & Markwardt, F. C. (1970). *Manual for the Peabody Individual Achievement Test*. Circle Pines, MN: American Guidance Service.
- Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.
- Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson & D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*. London, England: Chapman & Hall.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis* (2nd ed.). New York, NY: Chapman & Hall/CRC.
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733-760.

- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. New York, NY: Chapman & Hall/CRC.
- Habing, B., & Roussos, L. A. (2003). On the need for negative local item dependence. *Psychometrika*, 68, 435-451.
- Holzing, K. J., & Swineford, F. (1939). *A study in factor analysis: The stability of a bi-factor solution* (Supplementary Educational Monographs, No. 48). Chicago, IL: University of Chicago Press.
- Houts, C. R., & Edwards, M. C. (2013). The performance of local dependence measures with psychological data. *Applied Psychological Measurement*, 37, 541-562.
- Ip, E. H.-S. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, 66, 109-132.
- Lee, T., Cai, L., & Kuhfeld, M. (2016). A poor person's posterior predictive checking of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 206-220.
- Levy, R. (2006). *Posterior predictive model checking for multidimensionality in item response theory and Bayesian networks* (Unpublished doctoral dissertation). University of Maryland, College Park.
- Levy, R. (2011). Bayesian data-model fit assessment for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 18, 663-685.
- Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, 33, 519-537.
- Levy, R., & Svetina, D. (2011). A generalized dimensionality discrepancy measure for dimensionality assessment in multidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, 64, 208-232.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS—A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325-337.
- Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics*, 22, 1142-1160.
- Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.
- Rubin, D. B. (1987). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm. *Journal of the American Statistical Association*, 82, 543-546.
- Sinharay, S. (2016). Assessment of person fit using resampling-based approaches. *Journal of Educational Measurement*, 53, 63-85. doi:10.1111/jedm.12101
- Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, 30, 298-321.
- Sinharay, S., & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28, 237-247.
- Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6, 181-195.
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2002). Identification and evaluation of local item dependencies in the medical college admissions test. *Journal of Educational Measurement*, 39, 291-309.