Article

Applied Psychological Measurement
2017, Vol. 41(1) 3–16

DOI: 10.1177/0146621616669096
apm.sagepub.com
SAGE

Advancing the Bayesian Approach for Multidimensional Polytomous and Nominal IRT Models: Model Formulations and Fit Measures

Jinsong Chen[1]

Abstract

It is common to encounter polytomous and nominal responses with latent variables in social or behavior research, and a variety of polytomous and nominal item response theory (IRT) models are available for applied researchers across diverse settings. With its flexibility and scalability, the Bayesian approach using the Markov chain Monte Carlo (MCMC) method demonstrates its great advantages for polytomous and nominal IRT models. However, the potential of the Bayesian approach would not be fully realized without model formulations that can cover various models and effective fit measures for model assessment or criticism. This research first provided formulations for typical models that are representative of different modeling groups. Then, a series of discrepancy measures that can offer diagnostic information for model-data misfit were introduced. Simulation studies showed that the formulation worked as expected, and some of the fit measures were more useful than the others or across different situations.

Keywords

Bayesian, polytomous response, nominal response, model formulation, PPMC

Introduction

It is common to encounter polytomous or nominal responses with latent variables in social or behavior research, such as the uses of Likert-type or nominal scales for measuring psychological constructs or partial credit responses for assessing educational traits. Note that dichotomous responses can be subsumed as a special case. When polytomous or nominal models are formulated within the item response theory (IRT) framework, one can choose between the Bayesian approach based on the Markov chain Monte Carlo (MCMC; Gilks, Richardson, & Spiegelhalter, 1996) method or the conventional frequentist approach based on the full-information maximum likelihood (ML) or limited-information weighted least squares (WLS) methods. However, the two approaches are not necessarily competing with each other. Instead, the Bayesian approach

[1]Sun Yat-Sen University, Guangzhou, China

Corresponding Author:
Jinsong Chen, Department of Psychology, Sun Yat-Sen University, No. 135, Xingangxi Road, Guangzhou 510275, China.
Email: jinsong.chen@live.com

can be regarded as a viable alternative or important extension when the frequentist approach is less effective.

Specifically, there were a variety of polytomous and nominal IRT models, such as the graded response model (GRM; Samejima, 1969), rating scale model (RSM; Andrich, 1978), partial credit model (PCM; Masters, 1982), generalized partial credit model (GPCM; Muraki, 1992), generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), nominal response model (NRM; Bock, 1972), multiple-choice model (MCM; Thissen & Steinberg, 1984), and nested logit models (NLMs; Suh & Bolt, 2010), to name a few. The multidimensional versions of some models that were not available yet under the frequentist approach (e.g., GGUM, NRM, MCM, NLM) can be formulated and estimated using the Bayesian approach. Moreover, the Bayesian approach was preferred when confronting mixed response formats (Yao & Schwarz, 2006) or complex modeling such as integrating data from different studies with a hierarchical structure, multiple dimensions, and different sample/study characteristics (Huo et al., 2015). Finally, there usually existed many item parameters in polytomous or nominal models. Estimation of high dimensional models with a large number of parameters can be problematic under the frequentist approach, making the Bayesian approach the only way out.

For applied researchers, however, some pieces of the Bayesian puzzle are still missing before it can be accessible similar to the frequentist approach. Among others, the issues include the following: (a) the complexity of formulating various polytomous and nominal IRT models with the specifications of prior distributions and (b) the lack of overall fit measures such as the root mean square error of approximation (RMSEA; Hu & Bentler, 1999; Steiger & Lind, 1980) and standardized root mean residual (SRMR; Bentler, 1995; Hu & Bentler, 1999) that can summarize some aspects of model-data misfit in the absolute sense. For the first issue, one can build on the current Bayesian literature of polytomous IRT modeling (e.g., Arnold-Berkovits, 2002; Chen, Zhang, & Choi, 2015; de la Torre & Patz, 2005; Kieftenbeld & Natesan, 2012; Wirth & Edwards, 2007; Yao & Schwarz, 2006; Zhu & Stone, 2011). The variety of polytomous IRT models can be classified into two modeling groups (Reckase, 2009; Thissen & Steinberg, 1986): the difference-between-categories group represented by the GRM and the divide-by-total group represented by the GPCM. Nominal IRT models were best represented by the NRM, whose Bayesian estimation and formulation cannot be found in the current literature yet. The first purpose of this article is to cover all three representative models in a general formulation framework with the specifications of different prior distributions, which can be readily adapted by applied researchers for their purposes.

The second issue is of more concern, as the overall or global model-data fitness is critical to the validity of model-based measurement. Although there were other ways to construct fit measures under the Bayesian approach (e.g., Li, Cohen, Kim, & Cho, 2009; Yao & Schwarz, 2014), this research will focus on different discrepancy measures implemented via the posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Meng, 1994) method, which was found empirically useful in the IRT literature (e.g., Béguin & Glas, 2001; Karabatsos & Batchelder, 2003; Karabatsos & Sheu, 2004; Levy, Mislevy, & Sinharay, 2009; Li et al., 2009; Sinharay, Johnson, & Stern, 2006; Zhu & Stone, 2011). It is worth noting, however, that the PPMC method suffered concerns of the double use of data and being conservative when computing the p value (e.g., Bayarri & Berger, 1998; Robins, Van der Vaart, & Ventura, 2000), which necessitated more research into the topic. While various discrepancy measures had been empirically investigated in above studies, the focus was on item-level (e.g., item pair) rather than test-level fit information. Moreover, they were mainly targeted at models for dichotomous responses, except for Zhu and Stone's (2011), which adopted the unidimensional GRM. It is not straightforward to generalize fit measures useful for dichotomous responses to polytomous or nominal responses. Built on the current literature, this research proposes test-level

discrepancy measures for polytomous and nominal responses. Simulation studies are adopted to empirically evaluate if the fit measures are sensitive to model-data misfit across different settings. The simulation studies also help to verify the accuracy of parameter recovery under the formulation framework. In the rest of this research, the issue of model formulation is first addressed, followed by the introduction of the PPMC fit statistics. Then, design and results of the simulation studies are covered, with some discussions to end the research.

Theoretical Background

Model Formulation

A basic logic of most if not all polytomous and nominal IRT models is to split the M-category item into $M - 1$ binary subitems, each then can be formulated using the binary logistic models. For polytomous responses, the two modeling groups differ in how the item is split (Embretson & Reise, 2000; Thissen & Steinberg, 1986). Readers can refer to Reckase (2009) for more details of these two groups of modeling under the multidimensional settings, and below is a brief summary. In the difference-between-categories group, the item is split indirectly. If the probability of a response in response category m or above is denoted as $P^*(m)$, the probability of a response in category m is as follows:

$$P(m) = P^*(m) - P^*(m+1), \tag{1}$$

where $m = 0, \ldots, M - 1$ denote the response category (for simplicity but without loss of generality, we assume that all items have the same number of category, which can be relaxed in case needed). A representative model of this group is the GRM, which can be formulated using the normal ogive or logistic link function. With the logistic link, it is written as

$$P^*(m) = \frac{1}{\left(1 + \exp\left[-\alpha_{j(d)}\left(\theta_{i(d)} - \beta_{j(d),m}\right)\right]\right)}, \tag{2}$$

where $i = 1, \ldots, N$ denote the examinee; $d = 1, \ldots, D$ denote the dimension or domain of the test; $j(d) = 1, \ldots, J(d)$ denote the item in dimension d, and $J = \sum_{d=1}^{D} J(d)$ denote the total number of all items; $\theta_{i(d)}$ is the latent variable of examinee i on dimension or domain d; $\alpha_{j(d)}$ and $\beta_{j(d),m}$ are the slope and threshold parameters of item $j(d)$, respectively, with $\beta_{j(d),0} = -\infty$ and $\beta_{j(d),M} = +\infty$.

Alternatively, the normal ogives link function can be used, which is mathematically equivalent to factor analysis for categorical data (Takane & de Leeuw, 1987). As Thissen and Steinberg (1986) noted, it is necessary that the slope parameters be equal for all $P^*(m)$ in specific item given the normal ogive or logistic link function is used. Otherwise, one will get negative probability somewhere, and the modeling process would be failed. Under a similar argument, it is necessary that $\beta_{j(d),m} \geq \beta_{j(d),m'}$ for any $m > m'$, or equivalently the between-category threshold parameters be ordered, making the estimation of the threshold parameter tricky under the Bayesian approach. One solution is to let $\beta_{j(d),m+1} = \beta_{j(d),m} + \Delta\beta_{j(d),m}$ for $m > 0$, with $\Delta\beta_{j(d),m} \geq 0$ as the incremental threshold parameter, and estimate $\beta_{j(d),1}$ and $\Delta\beta_{j(d),m}$ instead (see Arnold-Berkovits, 2002; Chen et al., 2015, for more details).

In the divide-by-total group, the item is split directly with the probability of a response in either of two adjacent response categories. Mathematically, the probability of a response in specific category can be directly written as an exponential divided by the sum of all possible exponentials. A typical model in this group is the GPCM, which is

$$P(m) = \frac{\exp\left(\alpha_{j(d)} \sum_{s=0}^{m} \left(\theta_{i(d)} - \delta_{j(d),s}\right)\right)}{\sum_{t=0}^{M-1} \exp\left(\alpha_{j(d)} \sum_{s=0}^{t} \left(\theta_{i(d)} - \delta_{j(d),s}\right)\right)}, \tag{3}$$

with definitions similar above, except that $\delta_{j(d),m}$ is the step parameter with $\sum_{m=0}^{M-1} \delta_{j(d),m} = 0$ or $\delta_{j(d),0} = 0$. Furthermore, no ordinal relationship is required for the step parameters across categories. In practice however, similar ordinal relationship (i.e., $\delta_{j(d),m} \geq \delta_{j(d),m'}$ for any $m > m'$) is preferred for most items. Otherwise, the marginal frequencies of some response categories would be quite small compared with others in those items, adding to the difficulty of estimation of the corresponding item parameters (Muraki, 1992).

According to Maydeu-Olivares, Drasgow, and Mead (1994), the GRM and GPCM, which have the same number of item parameters, are almost interchangeable due to very close model fits. Although the GRM and GPCM were chosen to represent the two modeling groups in this research due to their similarity, one can easily obtain their constrained versions such as the RSM and PCM.

In its extreme form of the divide-by-total group, the slope parameters can be different across categories, which allows truly nominal responses and becomes the NRM eventually (Thissen & Steinberg, 1986), as follows:

$$P(m) = \frac{\exp\left(\alpha_{j(d),m}\theta_{i(d)} - \delta_{j(d),m}\right)}{\sum_{t=0}^{M-1} \exp\left(\alpha_{j(d),t}\theta_{i(d)} - \delta_{j(d),t}\right)}, \tag{4}$$

with the constraints $\sum_{m=0}^{M-1} \alpha_{j(d),m} = 0$ and $\sum_{m=0}^{M-1} \delta_{j(d),m} = 0$. Mathematically, the NRM subsumes the GPCM by allowing only one slope parameter per item, and accordingly its estimation is more challenging with more parameters. The NRM is the foundation of a series of multiple-choice models, such as the MCM (Thissen & Steinberg, 1984) and NLM (Suh & Bolt, 2010).

Model Estimation

In this research, the Bayesian approach via the MCMC estimation method is employed. From a Bayesian perspective, MCMC is a resampling-based simulation algorithm that iteratively resamples from the probability distributions such as the joint posterior and/or full conditional distributions based on a stochastic process of Markov chains (Gill, 2002). Readers can refer to Casella and George (1992); Chib and Greenberg (1995); Gelman, Carlin, and Stern (1995); and Gilks et al. (1996) for a general overview of MCMC. There are two flavors of MCMC that are widely used in practice: The Gibbs sampler (Casella & George, 1992) which is easier to implement but require the existence of full conditional distributions, and the Metropolis–Hastings (MH) algorithm (Chib & Greenberg, 1995) which only relies on the joint posterior distribution. In both flavors, the likelihood of the data matrix plays an important role.

Let $X_{ij(d)m} = 1$ if examinee i responds m on item j of dimension d, and 0 otherwise. Denote $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta\beta})$ or $\boldsymbol{\omega} = (\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\delta})$ as the collection of all model parameters for the GRM or GPCM and NRM, respectively, where $\boldsymbol{\theta} = \{\theta_{i(d)}\}$, $\boldsymbol{\Sigma}$ is the covariance matrix, $\boldsymbol{\alpha} = \{\alpha_{j(d)}\}$ for GRM and GPCM, or $\{\alpha_{j(d),m}\}$ for NRM, $\boldsymbol{\beta} = \{\beta_{j(d),1}\}$, $\boldsymbol{\Delta\beta} = \{\Delta\beta_{j(d),m}\}$, and $\boldsymbol{\delta} = \{\delta_{j(d),m}\}$. The likelihood of the data matrix $\mathbf{X} = \{X_{ij(d)m}\}$ is given as

$$L(\mathbf{X}|\boldsymbol{\omega}) = \prod_{i=1}^{N} \prod_{d=1}^{D} \prod_{j(d)=1}^{J(D)} \prod_{m=0}^{M-1} P(m)^{X_{ij(d)m}}, \tag{5}$$

with $P(m)$ in Equations 2, 3, or 4 for the GRM, GPCM, or NRM, respectively. With the likelihood function and the independent assumption of the person and item parameters, the joint posterior distributions of interest can be found as

$$P(\boldsymbol{\omega}|\mathbf{X}) \propto L(\mathbf{X}|\boldsymbol{\omega}) \times P(\boldsymbol{\theta}, \boldsymbol{\Sigma}|\mathbf{X}) \times P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta\beta} \text{ or } \boldsymbol{\delta}|\mathbf{X}). \tag{6}$$

However, the above joint distribution cannot be fully simplified into explicit form, so that samples can be drawn. Instead, one can decompose the joint distribution into different full conditional distributions for easier sampling as follows:

$$P(\boldsymbol{\theta}|\mathbf{X}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\boldsymbol{\theta}' \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}\right) L(\mathbf{X}|\boldsymbol{\omega}), \tag{7}$$

$$P(\boldsymbol{\Sigma}|\mathbf{X}, \boldsymbol{\theta}) = P(\boldsymbol{\Sigma}|\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}|\boldsymbol{\Sigma}) p(\boldsymbol{\Sigma}), \tag{8}$$

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Delta\beta}|\mathbf{X}) \propto L(\mathbf{X}|\boldsymbol{\omega}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) p(\boldsymbol{\Delta\beta}), \tag{9}$$

$$P(\boldsymbol{\alpha}, \delta|\mathbf{X}) \propto L(\mathbf{X}|\boldsymbol{\omega}) p(\boldsymbol{\alpha}) p(\boldsymbol{\delta}), \tag{10}$$

where Equation 9 or 10 is for the GRM or the GPCM and NRM, respectively.

For the person parameters, the following prior distributions were adopted:

$$\boldsymbol{\theta}_i|\boldsymbol{\Sigma} \sim MVN(0, \boldsymbol{\Sigma}), \tag{11}$$

$$\boldsymbol{\Sigma} \sim \textit{Inv-Wishart}_{v_0}\left(\Lambda_0^{-1}\right), \tag{12}$$

where *MVN* and *Inv-Wishart* are the multivariate normal and inverse Wishart distributions, respectively, $v_0 = D + 2$ and the diagonal and off-diagonal elements of Λ_0 were set to 1 and 0.5, respectively. The informativeness of the prior for $\sum$ is determined by the degrees of freedom (*df*) of the *Inv-Wishart*. A small *df* (i.e., $D+2$) is usually used to make the priors relatively uninformative. With the above, the full conditional distribution of $\sum$ can be rewritten as $\textit{Inv}-\textit{Wishart}_{v_N}(\Lambda_N^{-1})$, where $v_N = v_0 + N$ and $\Lambda_N = \Lambda_N + \sum_{i=1}^{N} \boldsymbol{\theta}_i \boldsymbol{\theta}_i'$, which can be directly sampled.

For the item parameters, the four-parameter Beta distribution,4-Beta(*a, b, c, d*), as the prior distributions, was adopted, which were found flexible to estimate IRT models under the Bayesian context (e.g., de la Torre & Patz, 2005; Huo et al., 2015). In the distribution, *a* and *b* are the shape parameters, and *c* and *d* are the location parameters to define the boundaries of the distribution. Accordingly, it is written as follows:

$$\alpha_{j(d)} \sim \text{4-Beta}(a_\alpha, b_\alpha, c_\alpha, d_\alpha), \tag{13}$$

$$\beta_{j(d),1} \sim \text{4-Beta}\left(a_{\beta 1}, b_{\beta 1}, c_{\beta 1}, d_{\beta 1}\right), \tag{14}$$

$$\Delta\beta_{j(d),m} \sim \text{4-Beta}\left(a_{\beta 1}, b_{\Delta\beta}, c_{\Delta\beta}, d_{\Delta\beta}\right), \tag{15}$$

$$\delta_{j(d),m} \sim \text{4-Beta}(a_\delta, b_\delta, c_\delta, d_\delta). \tag{16}$$

Hyperparameters can be chosen to make the priors flat, relatively informative, or to satisfy our purposes (e.g., to make sure the estimates are larger than zero). For flat priors, one can fix both shape parameters as one and adjust the location parameters to cover a wide range.

Posterior Predictive Model Checking

PPMC can assess the discrepancy between the observed data $\mathbf{X}$ and the replicated data $\mathbf{X}^{\text{rep}}$ from the model-based posterior predictive distribution (PPD) using some discrepancy measures. Specifically, the PPD is

$$P(\mathbf{X}^{\text{rep}}|\mathbf{X}) = \int_{\boldsymbol{\omega}} P(\mathbf{X}^{\text{rep}}|\mathbf{X}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathbf{X}) d\boldsymbol{\omega}, \tag{17}$$

where $p(\boldsymbol{\omega}|\mathbf{X})$ is the posterior distribution (i.e., probability density function) empirically obtained through the MCMC estimation described above. Discrepancy measures $D(\mathbf{X}, \boldsymbol{\omega})$ are defined to capture the relevant features of the data and/or the discrepancy between the data and the model. Large differences between the realized discrepancies $D(\mathbf{X}, \boldsymbol{\omega})$, based on the observed data, and the distribution of the predictive discrepancies $D(\mathbf{X}^{\text{rep}}, \boldsymbol{\omega})$, based on the PPD, are indicative of model-data misfit. A useful mechanism to summarize the above information is a posterior predictive p value (PPP value; Gelman et al., 1996; Meng, 1994), as defined below:

$$\text{PPP value} = P(D(\mathbf{X}^{\text{rep}}, \boldsymbol{\omega}) \geq D(\mathbf{X}, \boldsymbol{\omega})|\mathbf{X}). \tag{18}$$

When the model fits the data adequately, the realized discrepancies locate around the middle of the predictive discrepancies distributions and the PPP values will be near .5. When there is serious model misfit, the realized discrepancies would fall in either the upper or lower tail area of the distribution, and the PPP values will be close to 0 or unity. The above logic is sound given that the discrepancy measures are sensitive to model-data fit. In practice, however, it might not be easy to find measures that are always sensitive to misfit across various settings. In this research, several measures are proposed and will be evaluated for their sensitivity using simulation studies.

Extreme PPP values (close to 0 or unity) are suggestive of model failure, but they should not be viewed as a significance level in the usual sense. This is because PPMC is of more diagnostic nature and the traditional critical significance level or cutoff point might not be appropriate for PPMC. The PPP values do not refer to an explicit hypothesis test and are not necessarily uniformly distributed when the fitted model is in fact correct (Meng, 1994). Accordingly, this research is to treat PPMC and PPP values as a diagnostic evidence for, rather than a hypothesis testing of, model-data misfit. Moreover, this research focuses on summary statistics that can give signal of overall or test-level model-data fit, although many of these statistics can be used for item-level fit evaluation. Compared with the item level, test-level assessment is more critical to the validity of model-based measurement. The following summary discrepancy measures will be studied:

Total score distribution. Let t_v denote the possible total score any examinee can obtain, where v is ranged from 0 to $V = J(M - 1)$. The median PPP values of t_v can be evaluated. A statistics summarizing the discrepancy measure of overall model fit $D_t = \sum_{v=1}^{V} [t_v - E(t_v)]^2 / E(t_v)$, where $E(y)$ is the expectation of y, can also be evaluated. It is worth noting that it is not a cause of concern if D_t follows any specific distribution like chi-square because the PPMC method automatically provides the relevant reference distribution.

Square proportion of category. Let p_v denote the square of the proportion of each item category, excluding the last category of each item, where v is ranged from 1 to $V = J(M - 1)$. The median PPP values of p_v can be evaluated. Alternatively, a discrepancy measure that can summarize p_v is $D_p = \sum_{v=1}^{V} [p_v - E(p_v)]^2 / E(p_v)$.

Item-pair correlation. Let $r_{jj'}$ denote the Pearson correlation between any two items j and j' ($j' < j$), the median PPP values of $r_{jj'}$ or a discrepancy measure that can summarize $r_{jj'}$: $D_r = \sum_{j=1}^{J} \sum_{j'=1}^{j-1} [r_{jj'} - E(r_{jj'})]^2 / E(r_{jj'})$ can be evaluated.

Item category-pair (log) odds ratio. If we treat each item category as a subitem, there will be $V = J(M - 1)$ independent subitem categories. Let $o_{vv'}$ denote the odds ratio (OR) between any two independent item categories v and v' ($v' < v$), with $v = 1, \ldots, V$. Namely, $o_{vv'} = n_{11}n_{00}/n_{01}n_{10}$, where $n_{yy'}$ are the number of examinees who scored y on item v and y' on item v'. The median PPP values of $o_{vv'}$ or a discrepancy measure that can summarize $o_{vv'}$: $D_o = \sum_{v=1}^{V} \sum_{v'=1}^{v-1} [o_{vv'} - E(o_{vv'})]^2 / E(o_{vv'})$ can be evaluated. Alternatively, the log of odds ratio (LOR) between any two independent item categories v and v' ($l_{vv'}$) using $D_l = \sum_{v=1}^{V} \sum_{v'=1}^{v-1} [l_{vv'} - E(l_{vv'})]^2 / E(l_{vv'})$ can be evaluated.

It is worth noting that some statistics similar to the above discrepancy measures had been studied with PPMC under different contexts. Specifically, similar total score distribution was found sensitive to misfit of dichotomous IRT models (Béguin & Glas, 2001; Sinharay et al., 2006) but less effective for unidimensional GRM (Zhu & Stone, 2011). The item-pair odds ratios and Yen's Q_3 (Yen, 1993) statistics, which measures the correlations between item pairs after accounting for the latent ability level in examinees, were found sensitive to item-level misfit for both dichotomous IRT models (Levy et al., 2009; Sinharay et al., 2006) and unidimensional GRM (Zhu & Stone, 2011). But in the odds ratio, the polytomous responses were dichotomized, which could result in a loss of information. More importantly, test-level sensitivity of the last two statistics was yet to be found. Finally, note that among the above measures, D_t and D_r are only applicable to ordered polytomous responses, whereas D_p and D_o, and D_l can be applied to nominal responses.

Simulation Studies

Two simulation studies were adopted to evaluate the performance of the above model formulation and sensitivity of the PPMC discrepancy measures for model misfit: (a) the scenario of polytomous IRT models and (b) the scenario of nominal IRT models.

Design

Simulation 1: The scenario of polytomous models. Both the GRM and GPCM were used for data generation and model fitting, so that cross-fitting (i.e., using GRM to fit GPCM data and vice versa) was available to verify the interchangeability of the two models with the above formulation. The following simulation conditions were considered: $N = 500$ and 1,000; $J(d) = 6$ and 12; $M = 4$; the dimension $D = 3$ was used for data generation with $D = 2$ and 3 for model fitting. When the generating and fitting dimensions are different, model-data misfit occurs which can be used to evaluate the sensitivity of the fit statistics. The case of $D = 2$ represented the situation of model misfit, in which the items of the third dimension of data generation were evenly split into the two dimensions of model fitting. For each condition, 200 data sets were generated with $\boldsymbol{\theta}_i | \mathbf{R} \sim MVN(0, \mathbf{R})$, where $\mathbf{R}$ is a correlation matrix with off-diagonal elements $\rho = .3$ or .6. Item parameters $\alpha_{j(d)} \sim U(0.5, 3)$, $\beta_{j(d),1} \sim U(-3, -1)$, $\Delta\beta_{j(d),m} \sim U(0, 2)$, $\delta_{j(d),1} \sim U(-3, 0)$, $\delta_{j(d),2} \sim U(-1.5, 1.5)$, and $\delta_{j(d),3} \sim U(0, 3)$ were randomly generated for each item in each replication. For model fitting, the shape parameters of all item priors were fixed at 1, whereas the location parameters were (0, 10) for $\alpha_{j(d)}$ and $\Delta\beta_{j(d),m}$, or (−10, 10) for $\beta_{j(d),1}$ and $\delta_{j(d),m}$. For MCMC, the first 5,000 iterations were discarded as burn-in as the authors of the present study noticed that the Markov chains converged before 4,000 iterations in all cases. After that, 10,000

iterations were sampled to construct the posterior distribution. The estimation and simulation codes were custom written and implemented using the Ox program (see Doornik, 2007).

For model estimation, the recovery of both item and person (including the dimensional correlations) parameters based on the mean bias and root mean square error (RMSE) for the GRM and GPCM will be evaluated. For the person parameters, the estimation performance can also be evaluated when the generating and fitting models are different. For model assessment, the proportions of extreme PPP values for the above discrepancy measures will be evaluated. Due to the diagnostic nature of the PPMC method, extreme PPP values in two cases will be considered: (a) those that are less than .025 or greater than .975, and (b) those that are less than .05 or greater than .95. These two cases are akin to conducting two-tailed hypothesis tests with α = .05 and .10, respectively. The proportions of extreme PPP values are then the empirical Type I error rates when the true model is fitted, or the empirical statistical power when the incorrect model is fitted. Ideally, nominal or conservative (i.e., smaller than nominal) Type I error rates with high statistical powers would be expected.

Simulation 2: The scenario of nominal models. The NRM was used for data generation and model fitting. All simulation conditions were similar above except for the following: for data generation, $N = 1{,}000$, $\alpha_{j(d),m} \sim U(-2, 2)$, and $\delta_{j(d),m} \sim U(-2, 2)$; for model fitting, the location parameters of the item priors were $(-10, 10)$ for $\alpha_{j(d),m}$ and $\delta_{j(d),m}$, whereas the shape parameters were still fixed at 1. In a preliminary study, it was found that the signs of the $\alpha_{j(d),m}$ and θ estimates can be opposite to the true values if the initial values of all $\alpha_{j(d),m}$ were chosen randomly. However, if most (e.g., 70%) of the initial and true $\alpha_{j(d),m}$ were at the same sign, the issue was dismissed. In practice, content experts can be employed to specify the initial signs of the slope parameters before estimation. For model assessment, as D_t and D_r are only applicable to ordered polytomous responses, only D_p, D_o, and D_l were investigated in ways similar to Simulation 1.

Results

Simulation 1: The scenario of polytomous models. Recovery of the item and person parameters for either the GRM or GPCM is shown in Tables 1 (for item parameter recovery) and 2 (for person parameter recovery). As shown, the performance of both polytomous IRT models was similar, and both the accuracy and reliability of estimations tended to be better as the sample size, number of items, or dimensional correlation increases. This demonstrated the effectiveness of estimations using the model formulation under the Bayesian approach. The results of cross-fitting the two models are shown in Table 3 (person parameter recovery only). Not to our surprise, the results confirm the similarity between the two models.

For model assessment, it was found that difference in any fit statistics between the two models was trivial. Moreover, it was found that the fit statistics D_p was insensitive in any situation when model-data misfit occurred, which was dropped for further analysis. This was not unexpected due to the univariate nature of the measure (i.e., Square proportion of category). Table 4 presents the proportion of extreme PPP values when the correct GRM is used ($D = 3$ for both the generating and fitting models), which is similar to the Type I error. Except for D_o based on the odds ratio, all other measures were rather conservative for both α = .05 and .10. Similar situation was founded when the correct GPCM was used (Table 5). Moreover, the influence of factors such as sample size, magnitude of dimensional correlations, and number of items was limited. Accordingly, any of D_t, D_r, and D_l can provide diagnostic information to accept the correct model.

Compared with the Type I error about accepting the correct model however, researchers might be more concerned about the sensitivity of the fit statistics to model-data misfit (i.e., the

Table 1. Recovery of Item Parameters.

			GRM				GPCM			
			Mean bias		RMSE		Mean bias		RMSE	
N	*R*	*J*	Slope	Threshold	Slope	Threshold	Slope	Step	Slope	Step
500	.3	6	−0.073	−0.005	0.521	0.459	−0.068	0.007	0.461	0.470
		12	−0.044	−0.004	0.317	0.397	−0.048	−0.004	0.366	0.421
	.6	6	−0.072	−0.007	0.491	0.410	−0.051	−0.006	0.409	0.334
		12	−0.036	−0.002	0.311	0.369	−0.041	0.004	0.378	0.383
1,000	.3	6	−0.011	−0.001	0.289	0.263	0.028	0.001	0.267	0.218
		12	0.003	0.002	0.183	0.265	0.005	−0.002	0.277	0.230
	.6	6	0.005	0.001	0.197	0.241	−0.010	0.000	0.193	0.131
		12	−0.015	0.000	0.182	0.218	0.002	−0.003	0.197	0.173

Note. GRM = graded response model; GPCM = generalized partial credit model; RMSE = root mean square error.

Table 2. Recovery of Person Parameters.

			GRM				GPCM			
			Mean bias		RMSE		Mean bias		RMSE	
N	*R*	*J*	ρ	θ	ρ	θ	ρ	θ	ρ	θ
500	.3	6	−.006	0.041	.048	0.372	−.011	0.055	.049	0.382
		12	−.006	−0.045	.045	0.401	−.008	−0.048	.048	0.369
	.6	6	−.008	0.051	.038	0.284	−.004	−0.058	.049	0.374
		12	−.011	0.043	.034	0.273	−.008	0.056	.037	0.339
1,000	.3	6	−.004	−0.024	.033	0.362	−.008	0.052	.035	0.303
		12	−.002	0.031	.029	0.355	−.015	0.041	.04	0.279
	.6	6	−.005	0.025	.024	0.303	−.005	0.011	.034	0.268
		12	−.003	0.033	.028	0.279	−.001	0.026	.037	0.333

Note. Both the true and fitted models are the same. GRM = graded response model; GPCM = generalized partial credit model; RMSE = root mean square error.

Table 3. Recovery of Person Parameters for Different Generating and Fitting Models.

			GRM → GPCM				GPCM → GRM			
			Mean bias		RMSE		Mean bias		RMSE	
N	*R*	*J*	ρ	θ	ρ	θ	ρ	θ	ρ	θ
500	.3	6	.018	0.053	.054	0.535	−.016	−0.067	.059	0.585
		12	−.010	−0.051	.051	0.343	.013	−0.054	.051	0.488
	.6	6	.018	0.060	.043	0.502	.027	−0.056	.053	0.464
		12	−.011	−0.042	.041	0.385	.022	0.058	.039	0.386
1,000	.3	6	.013	0.035	.038	0.476	−.017	0.056	.041	0.382
		12	−.004	0.029	.035	0.373	.016	0.044	.039	0.293
	.6	6	.011	0.039	.033	0.360	−.020	−0.033	.035	0.375
		12	.009	−0.033	.031	0.291	.017	0.031	.034	0.273

Note. GRM → GPCM: The GRM is fitted to data generated from the GPCM; GPCM → GRM: The GPCM is fitted to data generated from the GRM. GRM = graded response model; GPCM = generalized partial credit model; RMSE = root mean square error.

Table 4. Proportion of Extreme PPP Values for Correct Models (GRM).

			α = .05				α = .10			
N	*R*	*J*	D_t	D_r	D_o	D_l	D_t	D_r	D_o	D_l
500	.3	6	.01	.00	.06	.03	.03	.00	.12	.08
		12	.02	.00	.09	.02	.03	.00	.11	.07
	.6	6	.02	.00	.05	.03	.03	.00	.12	.07
		12	.05	.00	.09	.02	.10	.00	.11	.08
1,000	.3	6	.00	.00	.03	.03	.01	.00	.05	.06
		12	.01	.00	.06	.01	.03	.00	.10	.02
	.6	6	.02	.00	.04	.02	.05	.00	.09	.05
		12	.03	.01	.03	.02	.09	.01	.07	.05

Note. Data are generated using GRM with $D = 3$; GRM is fitted with $D = 3$. PPP = posterior predictive *p* value; GRM = graded response model.

Table 5. Proportion of Extreme PPP Values for Correct Models (GPCM).

			α = .05				α = .10			
N	*R*	*J*	D_t	D_r	D_o	D_l	D_t	D_r	D_o	D_l
500	.3	6	.00	.00	.08	.01	.02	.00	.12	.02
		12	.00	.00	.07	.02	.04	.00	.12	.03
	.6	6	.02	.00	.06	.01	.04	.00	.10	.02
		12	.02	.00	.07	.01	.03	.00	.11	.01
1,000	.3	6	.00	.00	.05	.01	.01	.00	.08	.05
		12	.01	.00	.04	.01	.02	.00	.12	.05
	.6	6	.03	.00	.07	.02	.04	.00	.10	.06
		12	.00	.00	.05	.00	.00	.00	.10	.00

Note. Data are generated using GRM with $D = 3$; GPCM is fitted with $D = 3$. PPP = posterior predictive *p* value; GPCM = generalized partial credit model.

Table 6. Proportion of Extreme PPP Values for Incorrect Models (GRM).

			α = .05				α = .10			
N	*R*	*J*	D_t	D_r	D_o	D_l	D_t	D_r	D_o	D_l
500	.3	6	.96	1.00	.22	.76	.98	1.00	.27	.84
		12	.95	1.00	.31	1.00	.97	1.00	.41	1.00
	.6	6	.19	.67	.24	.18	.29	.74	.34	.30
		12	.25	.98	.36	.64	.38	.99	.48	.79
1,000	.3	6	1.00	1.00	.22	.97	1.00	1.00	.32	.99
		12	1.00	1.00	.49	1.00	1.00	1.00	.63	1.00
	.6	6	.56	.90	.36	.59	.65	.94	.45	.72
		12	.71	1.00	.65	.97	.78	1.00	.76	.97

Note. Data are generated using GRM with $D = 3$; GRM is fitted with $D = 2$. PPP = posterior predictive *p* value; GRM = graded response model.

statistical power to reject the incorrect model). Tables 6 and 7 present the proportion of extreme PPP values for detecting the incorrect models ($D = 3$ for model generation and $D = 3$ for model fitting). As shown, the item-pair correlation (i.e., D_r) had the highest power, while the OR (i.e.,

Table 7. Proportion of Extreme PPP Values for Incorrect Models (GPCM).

			$\alpha = .05$				$\alpha = .10$			
N	*R*	*J*	D_t	D_r	D_o	D_l	D_t	D_r	D_o	D_l
500	.3	6	.95	1.00	.28	.80	.99	1.00	.34	.89
		12	.90	1.00	.35	.75	.95	1.00	.45	.85
	.6	6	.24	.69	.32	.26	.35	.79	.39	.37
		12	.34	.87	.44	.59	.46	.94	.54	.73
1,000	.3	6	.91	1.00	.30	.85	.95	1.00	.45	.85
		12	.97	1.00	.41	.92	1.00	1.00	.57	.97
	.6	6	.53	.73	.36	.51	.54	.81	.41	.64
		12	.70	.94	.51	.91	.80	.85	.85	.95

Note. Data are generated using GRM with $D = 3$; GPCM is fitted with $D = 2$. PPP = posterior predictive *p* value; GPCM = generalized partial credit model; GRM = graded response model.

Table 8. Recovery of Item and Person Parameters for the NRM ($N = 1{,}000$).

		Item parameter				Person parameter			
		Mean bias		RMSE		Mean bias		RMSE	
R	*J*	Slope	Step	Slope	Step	ρ	θ	ρ	θ
.3	6	−0.007	−0.013	0.261	0.220	.000	0.036	.017	0.408
	12	−0.004	−0.012	0.237	0.211	.001	0.028	.034	0.305
.6	6	−0.004	−0.016	0.254	0.219	.000	0.027	−.010	0.413
	12	−0.002	−0.012	0.237	0.211	.001	0.021	−.006	0.296

Note. NRM = nominal response model; RMSE = root mean square error.

D_o) was the worst, with D_l being slightly better than D_t across all situations. As the sample size or number of items got larger, the power tended to increase. As the dimensional correlations got stronger, it became increasingly difficult (i.e., less powerful) to detect the incorrect model for most measures, except for D_o. In brief, three fit statistics were preferred, in order, to provide misfit diagnostic information for polytomous models: D_r, D_l, and D_t.

Simulation 2: The scenario of nominal models. Recovery of the item and person parameters for the NRM can be found in Table 8. In general, the performance of the NRM was close to that of the GRM or GPCM when $N = 1{,}000$. Similarly, the accuracy and reliability of estimations tended to be better as the number of items or dimensional correlation increases. Table 9 presents the proportion of extreme PPP values for D_o and D_l, when the correct or incorrect models are used (similarly, D_p is found insensitive to model-data misfit in any case and is dropped for further analysis). As shown, although the Type I errors were close to the nominal values and slightly conservative, the statistical powers were rather low (i.e., ~60% or lower in most cases) for both measures. One interesting finding was that the two measures can perform differently: D_l was more sensitive to smaller magnitude of dimensional correlations, whereas D_o was the opposite. This suggested that the two measures should be used together for better statistical power across different situations.

Table 9. Proportion of Extreme PPP Values for the NRM.

		Correct model (Type I)				Incorrect model (power)			
		$\alpha = .05$		$\alpha = .10$		$\alpha = .05$		$\alpha = .10$	
R	*J*	D_o	D_l	D_o	D_l	D_o	D_l	D_o	D_l
.3	6	.05	.04	.09	.07	.37	.47	.45	.60
	12	.01	.05	.03	.05	.47	.65	.64	.80
.6	6	.05	.06	.07	.07	.53	.27	.65	.42
	12	.02	.06	.06	.09	.82	.35	.87	.52

Note. PPP = posterior predictive *p* value; NRM = nominal response model.

Discussion

It is common to encounter polytomous and nominal responses with latent variables in social or behavior research, and a variety of polytomous and nominal IRT models are available for applied researchers across diverse settings. With its flexibility and scalability, the Bayesian approach using the MCMC method demonstrates its great advantages for polytomous and nominal IRT models. In addition, the MCMC method allows accommodation of most, if not all, IRT models in one setting using a specific programming platform. However, the potential of the Bayesian approach would not be fully realized for applied researchers without a general model formulation that can cover various polytomous and nominal models and effective fit measures for model assessment or criticism. Although this research only provided formulation for specific polytomous and nominal IRT models, the models are representative of different modeling groups that can accommodate most if not all polytomous and nominal models. In the future, the formulation can be extended to other special models such as the multidimensional versions of the MCM (Thissen & Steinberg, 1984), forced-choice model (Brown & Maydeu-Olivares, 2013), and NLM (Suh & Bolt, 2010).

For the proposed discrepancy measures, the simulation studies showed that some of them can provide useful diagnostic information regarding the correct and incorrect models. Specifically, the PPP values based on the item-pair correlations, item category-pair LOR, and total score distributions, in order of preference, were found sensitive to model-data misfit for polytomous responses. Moreover, the item category-pair OR and LOR can be used together for better statistical power in case of nominal responses. However, more effective or powerful fit measures are needed to address the misfit of nominal models. In addition, the Type I errors of most fit measures were smaller than the nominal values, which seems to be consistent with the conservative nature of the PPMC method. In light of this, one might consider further exploring fit measures in non-PPMC ways of model assessment (e.g., Li et al., 2009; Yao & Schwarz, 2014). Moreover, item-level version of these and other statistics can be investigated for model adjustment and improvement in future studies. Finally, the customized codes were inconvenient to use in practice, although they were excellent for simulation purpose. Applied researchers would prefer programming in more accessible and general-purpose Bayesian software like the BUGS (Lunn, Spiegelhalter, Thomas, & Best, 2009). Accordingly, development of program codes in such software, together with a didactic to illustrate the model formulations and fit measures with real-life examples, might be desirable.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by the Humanities and Social Sciences Research Grant from the Ministry of Education in China (Grant 14YJA880005). This research was also supported by the Humanities and Social Sciences Common Program from the Guangdong Province (Grant 2013WYXM0005).

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Arnold-Berkovits, I. (2002). *Structural modeling with ordered polytomous and continuous variables: A simulation study comparing full-information Bayesian estimation to correlation/covariance methods* (Doctoral dissertation). University of Maryland, College Park.

Bayarri, M. J., & Berger, J. O. (1998). Quantifying surprise in the data and model verification. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 6* (pp. 53-82). New York, NY: Oxford University Press.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*, 36-52. doi:10.1037/a0030641

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167-174.

Chen, J., Zhang, D., & Choi, J. (2015). Estimation of the latent mediated effect with ordinal data using the limited-information and Bayesian full-information approaches. *Behavior Research Methods*, *47*, 1260-1273. doi:10.3758/s13428-014-0526-3

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327-335.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295-311.

Doornik, J. A. (2007). *Object-oriented matrix programming using Ox* (3rd ed.). London, England: Timberlake Consultants Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Gelman, A., Carlin, J. B., & Stern, H. S. (1995). *Bayesian data analysis*. London, England: Chapman & Hall.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733-807.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. London, England: Chapman & Hall.

Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: Chapman & Hall/CRC.

Hu, L.-T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

Huo, Y., de la Torre, J., Mun, E.-Y., Kim, S.-Y., Ray, A., Jiao, Y., & White, H. (2015). A hierarchical multi-unidimensional IRT approach for analyzing sparse, multi-group data for integrative data analysis. *Psychometrika*, *80*, 834-855.

Karabatsos, G., & Batchelder, W. (2003). Markov chain estimation for test theory without an answer key. *Psychometrika*, *68*, 373-389.

Karabatsos, G., & Sheu, C.-F. (2004). Order-constrained Bayes inference for dichotomous models of unidimensional nonparametric IRT. *Applied Psychological Measurement*, *28*, 110-125.

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*, 399-419.

Levy, R., Mislevy, R. J., & Sinharay, S. (2009). Posterior predictive model checking for multidimensionality in item response theory. *Applied Psychological Measurement*, *33*, 519-537.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*, 353-373.

Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049-3067.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polytomous ordered data. *Applied Psychological Measurement*, *18*, 254-256.

Meng, X.-L. (1994). Posterior predictive p-values. *Annals of Statistics*, *22*, 1142-1160.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer-Verlag.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement*, *24*, 3-32.

Robins, J., Van der Vaart, A., & Ventura, V. (2000). Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, *95*, 1143-1156.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. New York, NY: Psychometric Society.

Sinharay, S., Johnson, D. M., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298-321.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically Based tests for the number of common factors*. Paper presented at the Annual Meeting of the Psychometric Society, Iowa City, IA.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454-473.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501-519.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*, 567-577.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*, 58-79.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement*, *30*, 469-492.

Yao, L., & Schwarz, R. (2014, April). *Comparison of methods in detecting the number of dimensions and item clusters for mixed format and mixed structured data using MCMC estimates*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Philadelphia, PA.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Zhu, X., & Stone, C. A. (2011). Assessing fit of unidimensional graded response models using Bayesian methods. *Journal of Educational Measurement*, *48*, 81-97.