Educational and Psychological Measurement
72(5) 774–799

DOI: 10.1177/0013164411434638
http://epm.sagepub.com

Bayesian Comparison of Alternative Graded Response Models for Performance Assessment Applications

Xiaowen Zhu[1] and Clement A. Stone[2]

Abstract

This study examined the relative effectiveness of Bayesian model comparison methods in selecting an appropriate graded response (GR) model for performance assessment applications. Three popular methods were considered: deviance information criterion (DIC), conditional predictive ordinate (CPO), and posterior predictive model checking (PPMC). Using these methods, several alternative GR models were compared with Samejima's unidimensional GR model, including the one-parameter GR model, the rating scale model, simple- and complex-structure two-dimensional GR models, and the GR model for testlets. Results from a simulation study indicated that these methods appeared to be equally accurate in selecting the preferred model. However, CPO and PPMC can be used to compare models at the item level, and PPMC can also be used to compare both the relative and absolute fit of different models.

Keywords

Bayesian model comparison, DIC, CPO, posterior predictive model checking, graded response models, multidimensionality, local independence

[1]Xi'an Jiaotong University, Xi'an, China
[2]University of Pittsburgh, Pittsburgh, PA, USA

Corresponding author:
Xiaowen Zhu, Department of Sociology & Institute for Empirical Social Science Research (IESSR), Xi'an Jiaotong University, Xi'an, Shaanxi, 710049, China
Email: xwzhu@mail.xjtu.edu.cn

With the increased popularity of including performance-based items in large-scale assessments, standard unidimensional polytomous item response theory (IRT) models are commonly used to analyze performance assessment data. However, the underlying assumptions (e.g., unidimensionality and local item independence) may be violated for performance assessment responses (Lane & Stone, 2006), and thus, more complex polytomous models may be needed for analysis. For example, a multidimensional polytomous model may be more appropriate for responses exhibiting multidimensionality, or a polytomous model for testlets may be more suitable for assessments that involve a subset of items with a common stimulus. To choose an appropriate model for a particular testing application, model comparison techniques can be employed.

For more complex IRT models, it has become increasingly common to estimate models using Bayesian methods with Markov Chain Monte Carlo (MCMC) algorithms. For these methods, different types of Bayesian approaches have been discussed for comparing competing models: (a) Bayesian information-based criteria, such as deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) and (b) indices based on Bayes factors (BF; Kass & Raftery, 1995), such as the pseudo-Bayes factor (PsBF; Geisser & Eddy, 1979; Gelfand, Dey, & Chang, 1992) and the conditional predictive ordinate (CPO; Kim & Bolt, 2007). Also, the posterior predictive model checking (PPMC) method can be used for model comparisons, though it is essentially a Bayesian tool for evaluating model data-fit (e.g., Béguin & Glas, 2001; Li, Bolt, & Fu, 2006; Sinharay, 2005).

Research comparing different model comparison methods in IRT applications has been limited. Li et al. (2006) investigated the performance of three Bayesian methods (DIC, PsBF, and PPMC) in selecting a testlet model for dichotomous item responses. They found that PsBF and PPMC were effective in identifying the data-generating testlet models, but DIC tended to favor more complex testlet models. Kang and Cohen (2007) examined the accuracy of the G^2 likelihood test statistic, Akaike's information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1978), DIC, and PsBF indices in selecting the most appropriate dichotomous IRT model. The PsBF index was found to be the more accurate index. The DIC index tended to select a more complex model, and conversely, the BIC index tended to select a simpler model. Kang, Cohen, and Sung (2009) later compared the performance of the AIC, BIC, DIC, and PsBF indices with respect to the selection of unidimensional polytomous IRT models. They found that BIC was accurate and consistent in selecting the true or simulated polytomous model. Performance of AIC was similar to BIC whereas DIC and PsBF were less accurate in some conditions when the true model was the graded response (GR) model.

The above findings indicate that model comparison methods perform inconsistently in Bayesian IRT applications. Their performance has been dependent on specific conditions as well as specific models that were compared. The purpose of this study was to extend previous research specifically to performance assessment applications and to investigate the effectiveness of different Bayesian methods in selecting a model

from a set of alternative models that are applicable to performance assessments. The set of models included Samejima's (1969) unidimensional two-parameter GR (2P-GR) model, a one-parameter version of the GR model (1P-GR), the rating scale (RS) model (Muraki, 1990), the simple- and complex-structure two-dimensional GR models (De Ayala, 1994), and the GR model for testlets (Wang, Bradlow, & Wainer, 2002). All models were estimated in WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) using MCMC methods, and compared using three Bayesian model comparison approaches: the DIC index, the CPO index, and the PPMC method.

Model Comparison Criteria

Deviance Information Criterion

DIC (Spiegelhalter et al., 2002) is an information-based index similar to other information criteria indices in that it weights both model fit and model complexity in identifying a preferred model. The DIC index is widely used with Bayesian estimation with MCMC methods and defined as

$$DIC = \bar{D}(\boldsymbol{\eta}) + p_D. \tag{1}$$

The first term, $\bar{D}(\boldsymbol{\eta})$, is the posterior mean of the deviance between the data and a model and defined as

$$\bar{D}(\boldsymbol{\eta}) = E_{\boldsymbol{\eta}|\mathbf{y}}[D(\boldsymbol{\eta})] = E_{\eta|y}[-2\log p(\mathbf{y}|\boldsymbol{\eta})], \tag{2}$$

where $\mathbf{y}$ represents the response data, $\boldsymbol{\eta}$ denotes the model parameters, and $p(\mathbf{y}|\boldsymbol{\eta})$ is the likelihood function. The second term in Equation (1), p_D, represents the effective number of parameters in the model or model complexity. It is defined as the difference between the posterior mean of the deviance and the deviance at the posterior mean of model parameters $\hat{\eta}$:

$$p_D = \bar{D}(\eta) - D(\hat{\eta}). \tag{3}$$

Note that DIC is a Bayesian generalization (Speigelhalter et al., 2002) of AIC and is related to BIC. DIC was developed since MCMC estimation uses prior information and the actual number of parameters cannot be clearly identified as required for the AIC and BIC indices. As for AIC and BIC, smaller values of DIC indicate better fit. DIC was the focus of this study since it is more commonly used for comparing models estimated with Bayesian methods and is more readily accessible to researchers using WinBUGS.

Conditional Predictive Ordinate

Bayes factors (BF; Kass & Raftery, 1995) have traditionally been used with Bayesian methods to compare competing models based on the observed data. The BF for

comparing Model 1 (M_1) with Model 2 (M_2) is defined by the posterior odds of M_1 to M_2 divided by the prior odds of M_1 to M_2. Using Bayes theorem, the BF reduces to the ratio of marginal likelihoods of the data (y) under each model:

$$BF = \frac{P(M_1|y)/P(M_2|y)}{P(M_1)/P(M_2)} = \frac{P(y|M_1)}{P(y|M_2)}. \tag{4}$$

A BF larger than 1 supports selection of M_1 and a value less than 1 supports selection of M_2. Research has shown that there are several issues with using this method in practice. For instance, the calculation of BF from MCMC output becomes difficult for complex models, and it is not well defined for improper priors (Li et al., 2006). As a result, PsBF (Geisser & Eddy, 1979; Gelfand et al., 1992) and CPO (Kim & Bolt, 2007) indices have been discussed as surrogates for BF.

Both PsBF and CPO indices require calculation of cross-validation predictive densities. Let $\mathbf{y}_{(r),\,\text{obs}}$ denote the set of observations $\mathbf{y}_{\text{obs}}$ with the r^{th} observation omitted, and let $\boldsymbol{\eta}$ denote all the parameters under the assumed model. The cross-validation predictive density is defined as

$$f\left(y_r|\mathbf{y}_{(r)}\right) = \int f\left(y_r|\boldsymbol{\eta}, \mathbf{y}_{(r)}\right) f\left(\boldsymbol{\eta}|\mathbf{y}_{(r)}\right) d\boldsymbol{\eta}. \tag{5}$$

The density $f(y_r|\mathbf{y}_{(r)})$ indicates the values of y_r that are likely when the model is estimated from all observations except y_r. In the context of item response data, y_r represents a single examinee's response to an individual item. The product of $f(y_r|\mathbf{y}_{(r)})$ across all observations can be used as an estimate of the marginal likelihood in Equation (4). The PsBF of Model 1 against Model 2 is defined as

$$PsBF = \frac{\prod_{r=1}^{R} f\left(y_{r,\,\text{obs}}|\mathbf{y}_{(r),\,\text{obs}}, M_1\right)}{\prod_{r=1}^{R} f\left(y_{r,\,\text{obs}}|\mathbf{y}_{(r),\,\text{obs}}, M_2\right)}, \tag{6}$$

where R denotes the total number of item responses from all examinees. When comparing the models at the item level, R equals the number of examinees N. When comparing the models at the test level, R equals the number of the responses for all examinees to all items (i.e., $R = N \times I$, where I is the total number of items). As for the BF index, a PsBF greater than 1 supports selection of Model 1 and a value less than 1 supports selection of Model 2.

The CPO index for a model can be defined as

$$CPO_M = \prod_{r=1}^{R} f\left(y_{r,\,\text{obs}}|\mathbf{y}_{(r),\,\text{obs}}, M\right). \tag{7}$$

The model with larger CPO value is preferred. It is easy to see that the same conclusions will be obtained using either CPO or PsBF indices, so only the CPO index was used in the current study.

Posterior Predictive Model Checking

PPMC (Rubin, 1984) is a flexible Bayesian model-checking tool that has recently proved useful in assessing different aspects of fit for IRT models (e.g., Fu, Bolt, & Li, 2005; Levy, Mislevy, & Sinharay, 2009; Sinharay, 2005, 2006; Sinharay, Johnson, & Stern, 2006; Zhu & Stone, 2011). PPMC involves simulating data using posterior distributions for model parameters and comparing features of simulated data against observed data. A discrepancy measure computed on both observed and simulated data is used to evaluate whether differences reflect potential misfit of a model (Gelman, Carlin, Stern, & Rubin, 2003). The comparison of realized (or observed) and posterior predictive (or simulated) values for the discrepancy measure can be performed using graphical displays as well as by computing posterior predictive p (PPP) values. PPP values provide numerical summaries of model fit. They reflect the relative occurrence of a value for a discrepancy measure $D(\mathbf{y}, \boldsymbol{\eta})$ based on the observed data $\mathbf{y}$ in the distribution of discrepancy values from replicated values $D(\mathbf{y}^{\text{rep}}, \boldsymbol{\eta})$:

$$PPP = P(D(\mathbf{y}^{\text{rep}}, \boldsymbol{\eta}) \geq D(\mathbf{y}, \boldsymbol{\eta}) | \mathbf{y}) = \iint_{D(\mathbf{y}^{\text{rep}}, \boldsymbol{\eta}) \geq D(\mathbf{y}, \boldsymbol{\eta})} p(\mathbf{y}^{\text{rep}} | \boldsymbol{\eta}) p(\boldsymbol{\eta} | \mathbf{y}) d\mathbf{y}^{\text{rep}} d\boldsymbol{\eta}, \quad (8)$$

where $\mathbf{y}$ represents the response data, $\mathbf{y}^{\text{rep}}$ represents the replicated data based on the model, and $\boldsymbol{\eta}$ is a vector of model parameters. PPP values near 0.5 indicate that there is no systematic difference between the realized and predictive discrepancies, and thus indicate adequate fit of the model. PPP values near 0 or 1 (e.g., values $<.05$ or $>.95$) suggest conversely that the realized discrepancies are inconsistent with the posterior predictive discrepancy values, and thus indicate inadequate model-data-fit (e.g., Gelman et al., 2003; Levy et al., 2009; Sinharay, 2005).

PPMC has been found useful in model comparisons when Bayesian estimation is used (e.g., Béguin & Glas, 2001; Li et al., 2006; Sinharay, 2005). The relative fit of a set of alternative candidate models can be evaluated by comparing the numbers of extreme PPP values across items (or item pairs). A model with fewer numbers of items (or item pairs) with extreme PPP values is considered to fit the data better than an alternative model with larger numbers of items (or item pairs) with extreme PPP values.

Selecting appropriate discrepancy measures is an important consideration in applications of the PPMC method. Discrepancy measures should be chosen to reflect sources of potential misfit most relevant to a testing application. In this study, the proposed measures included four item-level measures: *the item score distribution*, *the item–total score correlation*, *Yen's Q_1 statistic* (Yen, 1981), and *the pseudo-count fit* (e.g., Stone, 2000); and two pairwise measures: *global odds ratios* (*global ORs*;

Agresti, 2002), and *Yen's Q_3 statistic* (Yen, 1993). These particular measures were selected from a broader set based on their usefulness to detect IRT model-data-fit in previous testing applications (cf., Fu et al., 2005; Levy et al., 2009; Sinharay, 2005, 2006; Sinharay et al., 2006; Zhu & Stone, 2011).

1. Item-level measures. The *item score distribution* is an intuitive measure for assessing item-level fit. It represents the number of examinees responding to each response category for each item. The difference between observed and posterior predictive item score distributions can be summarized using a goodness-of-fit statistic (χ_j^2). For polytomous items, this statistic can be defined as

$$\chi_j^2 = \sum_{k=0}^{M_j} \frac{\left[O_{jk} - E_{jk}\right]^2}{E_{jk}}, \tag{9}$$

where M_j is the highest score on Item j, and O_{jk} and E_{jk} are the observed and posterior predicted number of examinees scoring in response category k on Item j, respectively. E_{jk} is calculated by summing the probabilities of responding to category k on Item j across all N examinees: $E_{jk} = \sum_{i=1}^{N} p_{ijk}(\theta_i)$.

The *item–total score correlation* measure is the correlation between examinees' total test scores and their scores on a particular item. This measure reflects item discrimination and therefore should be sensitive to any violations of the equal discrimination assumption. Sinharay et al. (2006) found that this correlation was effective in detecting misfit of 1PL models to data based on 2PL or 3PL models. Therefore, it was expected that this measure would be also useful in discriminating between the 1P-GR and 2P-GR models to be compared in this study. Pearson correlations were used to estimate associations between the five-category items and total test scores.

A number of other statistics have been proposed to assess the fit of IRT models at the item level (e.g., Orlando & Thissen, 2000; Stone, 2000; Yen, 1981). All these statistics compare observed response score distributions and model predictions at discrete levels of ability for each item. These item-fit statistics performed well for detecting item misfit in the frequentist framework, and several were also found effective with PPMC (e.g., Sinharay, 2006; Zhu & Stone, 2011). In this study, two of these statistics were used with PPMC: *Yen's Q_1* (Yen, 1981) and the *pseudo-count fit* statistic (e.g., Stone, 2000). The former is a traditional item-fit index and the latter is an alternative index developed specifically for performance assessment applications.

Yen's Q_1 item-fit statistic is a goodness-of-fit statistic defined as

$$\chi^2 = \sum_{j=1}^{10} \sum_{k=1}^{K} N_j \frac{\left(O_{jk} - E_{jk}\right)^2}{E_{jk}}, \tag{10}$$

where N_j is the number of examinees within ability subgroup j, O_{jk} and E_{jk} are the observed and predicted proportion of responses to category k for ability subgroup j, respectively. In *Yen's statistic*, examinees are divided into 10 ability subgroups of

approximately equal size after they are rank-ordered by their ability estimates. The expected proportion to a response category for a subgroup is the mean of the probabilities of responses to that category for all the examinees in that subgroup.

The *pseudo-count fit statistic* is a goodness-of-fit statistic developed to account for the imprecision in ability estimation for short tests, such as performance assessments. Whereas *Yen's* Q_1 cross-classifies examinees into only one cell of the item-fit table based on each item response and point estimate of ability, the *pseudo-count fit statistic* assigns each examinee to multiple ability groups based on posterior expectations (posterior probabilities for each discrete ability level). Summing the posterior expectations across all examinees provides a pseudo-observed score distribution. Model-based predictions for score categories at each ability level are calculated in the same way as in *Yen's* Q_1 statistic.

It may be worthy to note that a difference between these statistics and the *item-score distribution* index is that the observed and model-based predictions are compared at different discrete levels of the ability parameter in IRT models. Thus, summations in the equations are across response categories and discrete ability levels.

2. *Pairwise measures.* *Yen's* Q_3 index (Yen, 1981) and an *OR* measure were used as possible pairwise measures in this study. These measures are classic local-dependence indices and reflect the association between responses to item pairs. *Yen's* Q_3 statistic measures the correlations between pairs of items after accounting for the latent ability. For Items j and j', Q_3 is defined as the correlation of deviation scores across all examinees: $Q_{3jj'} = r\left(d_j, d_{j'}\right)$, where d_j is the deviation between observed and expected performance on Item j. Alternatively, Chen and Thissen (1997) used OR to evaluate local dependence for dichotomous items. The OR for dichotomous item pairs (j and j*) are computed from 2×2 tables by $n_{00}n_{11}/n_{01}n_{10}$, where n_{pq} is the observed number of examinees having response p (0 or 1) on Item j and response q (0 or 1) on Item j*.

In contrast with dichotomously scored items, multiple ORs can be computed with polytomous items since the contingency table is $R \times C$ ($R > 2$ and $C > 2$). For this study, a *global OR* (Agresti, 2002) was used as a possible discrepancy measure. For any two items, the $R \times C$ contingency table may be reduced to a 2×2 contingency table by dichotomizing the response categories of each item. The *global OR* was then defined as the cross-ratio of this pooled 2×2 table. In this study, the dichotomization was based on score rubrics often used with performance assessments. For items with 5 response categories (0-4), Categories 3 and 4 were treated as ''correct'' responses, and Categories 0, 1, and 2 were treated as ''incorrect'' responses.

Both *Yen's* Q_3 and *OR* measures have been found to be effective in detecting multidimensionality and local dependence among item responses with PPMC (e.g., Levy et al., 2009; Sinharay et al., 2006; Zhu & Stone, 2011). Therefore, they were expected to be useful in comparing unidimensional models with multidimensional IRT models or testlet models considered in this study. Although Q_3 has been found to be more effective than the *OR* measure when evaluating the fit of a single IRT model with PPMC (e.g., Levy et al., 2009; Zhu & Stone, 2011), both measures were used in the

present study to evaluate their respective usefulness in the context of a model comparison approach.

Alternative Graded Response Models for Performance Assessments

Samejima's (1969) 2P-GR model is a commonly used polytomous model for performance assessment applications. For Item j with $(m_j + 1)$ response categories (0, 1, 2, . . ., m_j), the probability that an examinee receives a category score x ($x = 1, 2, \ldots, m_j$) or higher $(P^*_{ijx}(\theta_i))$ is modeled by the logistic deviate e^Z, where $z = \lfloor Da_j(\theta_i - b_{jx}) \rfloor$, D is the scaling constant (1.7 or 1), θ_i is examinee ability, a_j is the discrimination (slope) parameter for Item j, and b_{jx} is a threshold parameter for Category x of Item j. The probability of a particular response in Category x for an examinee on Item j $(P_{ijx}(\theta_i))$ is then defined as the difference between the cumulative probabilities for two adjacent categories: $P^*_{ij(x-1)}(\theta_i)$ and $P^*_{ijx}(\theta_i)$, with two constraints $P^*_{ij0}(\theta_i) = 1$ and $P^*_{ij(m_i+1)}(\theta_i) = 0$.

For the standard 2P-GR model, one slope parameter is estimated for each item. Under the assumption that all items have the same discrimination, the 2P-GR model is reduced to a 1P-GR model with $z = \lfloor Da(\theta_i - b_{jx}) \rfloor$. Therefore, the 1P model is a restricted version of the standard GR model with a common slope parameter.

Another restricted case of the 2P-GR model is Muraki's (1990) RS model. In this model, $z = \lfloor Da_j(\theta_i - (b_j - c_x)) \rfloor$ and the threshold parameters (b_{jx}) in the 2P-GR model are partitioned into two terms: a location parameter (b_j) for each item, and one set of category threshold parameters (c_x) that applies to all items. The RS model is a restricted version of the 2P-GR model since category threshold parameters are assumed to be equal across all items in the RS model, whereas they are free to vary across items in the 2P-GR model. As a result, the number of parameters in the RS model is reduced greatly as compared with the standard 2P-GR model. The RS model was originally developed for analyzing responses to items with a rating-scale type response format. However, Lane and Stone (2006) pointed out that the RS model may be appropriate for performance assessments. When a general rubric is used as the basis for developing specific item rubrics, the response scales and the differences between score levels may be the same across the set of items.

The above models are appropriate for analyzing item responses that are assumed to be determined by one latent trait (i.e., unidimensional). When a performance assessment is designed to measure more than one ability, the item responses exhibit a multidimensional structure. De Ayala (1994) discussed a multidimensional version of the GR model, where $z = \left[D \sum_h a_{jh}(\theta_h - b_{jx}) \right]$ and θ_h is the ability level on dimension h, a_{jh} is the discrimination (slope) parameter of Item j on dimension h, and b_{jx} is the threshold parameter for Category x of Item j.

Finally, in performance assessments, a single stimulus (e.g., passage) may be used with several items (Yen, 1993). The responses to these items are therefore likely to

Table 1. Design and Conditions in this Simulation Study

Data generating model (*Mg*)	Data analysis model (*Ma*)	Condition Number
Two-parameter unidimensional GR (2P-GR)	2P-GR vs. 1P-GR vs. RS	1
Two-dimensional simple-structure GR (2D simple-GR)	2P-GR vs. 2D simple-GR	2
Two-dimensional complex-structure GR (2D complex-GR)	2P-GR vs. 2D complex-GR	3
Testlet GR	2P-GR vs. testlet GR	4

Note. GR = graded response; 1P-GR = 1-parameter unidimensional GR model; RS = rating scale model.

be locally dependent, and standard IRT models may be inappropriate. A modified GR model for testlets was proposed by Wang, Bradlow, and Wainer (2002), where $z = \left\lfloor Da_j\left(\theta_i - b_{jx} - \gamma_{id(j)}\right)\right\rfloor$. In this model, a random testlet effect ($\gamma_{id(j)}$) is introduced to reflect an interaction for Person i with testlet $d(j)$. Ability θ_i is typically assumed to have a $N(0, 1)$ distribution, and $\gamma_{id(j)}$ is assumed to be distributed as $N(0, \sigma^2_{d(j)})$. The variance of $\gamma_{id(j)}$ varies across testlets, and its value indicates the amount of local dependence in each testlet. As $\sigma^2_{d(j)}$ increases, the amount of local dependence increases. When $\sigma^2_{d(j)} = 0$, the items within the testlet are conditionally independent.

Simulation Study

Design of the Simulation Study

A simulation study was conducted to evaluate the use of Bayesian model comparison methods for selecting a model among alternative or competing GR models for performance assessment applications. The set of alternative models reflected sources of potential misfit for standard GR models and reflected models that may be theoretically more appropriate for some types of performance assessments.

Table 1 presents the design and specific conditions used in this simulation, and Table 2 presents item parameters for each GR model. For each condition, 20 response data sets were generated based on the data-generating model (*Mg*) with each data set containing responses for 2,000 simulated examinees to 15 polytomous items with 5 response categories. The test length and the number of response categories were fixed at typical values in performance assessment applications. Note that performance assessments are comprised commonly of forms with fewer items than multiple-choice type tests (Lane & Stone, 2006). A large sample size of 2,000 was used to ensure that model parameters were estimated precisely, and model selection results would not be affected by any inaccuracy in model parameter estimation.

Condition 1 was used to evaluate the effectiveness of the three model comparison methods in discriminating between the 2P-GR, 1P-GR, and RS models for the unidimensional responses simulated under the 2P-GR model. As shown in Table 2, the

Table 2. Item Parameters for Alternative Graded Response (GR) Models Under Conditions 1 to 4

	2P-GR and testlet GR					2D simple-GR		2D complex-GR	
Item	a	b1	b2	b3	b4	a1	a2	a1	a2
1	1.0	−2.0	−1.0	0.0	1.0	1.0	0.0	1.0	0.5
2	1.0	−1.5	−0.5	0.5	1.5	1.7	0.0	1.0	0.5
3	1.0	−1.0	0.0	1.0	2.0	2.4	0.0	1.0	0.5
4	1.0	−3.0	−1.5	−0.5	1.0	1.0	0.0	1.0	0.5
5	1.0	−1.0	0.5	1.5	3.0	1.7	0.0	1.0	0.5
6	1.7	−2.0	−1.0	0.0	1.0	2.4	0.0	1.7	0.0
7	1.7	−1.5	−0.5	0.5	1.5	1.0	0.0	1.7	0.0
8	1.7	−1.0	0.0	1.0	2.0	1.7	0.0	1.7	0.0
9	1.7	−3.0	−1.5	−0.5	1.0	0.0	2.4	1.7	0.0
10	1.7	−1.0	0.5	1.5	3.0	0.0	1.0	1.7	0.0
11	2.4	−2.0	−1.0	0.0	1.0	0.0	1.7	2.4	0.0
12	2.4	−1.5	−0.5	0.5	1.5	0.0	2.4	2.4	0.0
13	2.4	−1.0	0.0	1.0	2.0	0.0	1.0	2.4	0.0
14	2.4	−3.0	−1.5	−0.5	1.0	0.0	1.7	2.4	0.0
15	2.4	−1.0	0.5	1.5	3.0	0.0	2.4	2.4	0.0

configuration of item parameters for the 2P-GR model involved a combination of three slope parameters: 1.0, 1.7, and 2.4 (reflecting low, average, and high discrimination) and five threshold parameter configurations (reflecting different difficulty levels): (a) −2.0, −1.0, 0.0, 1.0; (b) −1.5, −0.5, 0.5, 1.5; (c) −1.0, 0.0, 1.0, 2.0; (d) −3.0, −1.5, −0.5, 1.0; (e) −1.0, 0.5, 1.5, 3.0. Ability parameters were randomly simulated from a $N(0, 1)$ distribution.

In Condition 2, the simulated test measured two dimensions but each item measured only one dimension (i.e., simple-structure case). Specifically, the first eight items were designed to measure one dimension, and the remaining seven items measured the second dimension. As shown in Table 2, the slope parameters on Dimension 1 (a_1) ranged from 1.0 to 2.4 for Items 1 to 8, and 0s for Items 9 to 15. The slope parameters on Dimension 2 (a_2) were 0 for Items 1 to 8 and ranged from 1.0 to 2.4 for Items 9 to 15. The threshold parameters for these 15 items were the same as for the 2P-GR model. Ability parameters on two dimensions were randomly selected from a bivariate normal distribution (0, 1) with a fixed correlation of 0.60, a correlation that may represent a value typical for large-scale performance assessments (cf., Lane, Stone, Ankenmann, & Liu, 1995). The responses were simulated based on the multidimensional GR model, and the three methods were used to compare the fit of the unidimensional 2P-GR model and the two-dimensional simple-structure (2D simple) GR model.

For Condition 3, two-dimensional complex-structure (2D complex case) responses were simulated to reflect a performance assessment that not only measures a

dominant ability (e.g., math word problems) but also consists of items that measure a nuisance or construct-irrelevant dimension (e.g., reading). All 15 items measured the dominant dimension, but the first 5 items were designed to also measure a nuisance dimension (see Table 2). The degree to which item performance depended on the nuisance ability was captured by the ratio of the slope (a_1) for the dominant ability to the slope (a_2) for the nuisance ability. The ratio of a_2 to a_1 was set to 0.5 for the first five items. The thresholds (b_1-b_4) and slope parameters for the dominant ability (a_1) were the same as for the 2P-GR model (see Table 2). The values of a_2 were 0.5 for Items 1 to 5 and 0.0 for other items. Ability parameters for two dimensions were randomly selected from a bivariate normal distribution (0, 1) with a correlation of 0.3. A low level of correlation was used because the second dimension was assumed to be either a nuisance dimension or construct-irrelevant dimension. The responses were simulated based on the multidimensional GR model, and the three methods were used to compare the fit of the unidimensional 2P-GR model and the 2D complex GR model.

In Condition 4, responses to a test consisting of a testlet were generated under the modified GR model for testlets. As shown in Table 2, item parameters for this testlet GR model were the same as that for the 2P-GR model, except one testlet was simulated (Items 6, 7, and 8). Ability parameters were randomly selected from $N(0, 1)$, and the testlet effect $\gamma_{jd(i)}$ was randomly selected from a $N(0, \sigma^2_{d(i)})$ for the items in the testlet. The variance of the testlet effect $\sigma^2_{d(i)}$ was specified at 1.0 reflecting a moderate degree of dependence among the testlet items (Bradlow, Wainer, & Wang, 1999; Li et al., 2006; Wang et al., 2002).

For each condition, 20 data sets of 2,000 item responses were generated using SAS based on the *Mg* model described in Table 1. Using the specified model parameters (i.e., item and ability), the probabilities of each simulated examinee responding to each item response category were calculated, and these probabilities were used to obtain simulated discrete item responses (see Zhu, 2009, for complete results). For each of the generated data sets within each condition, different analysis models (*Ma*) were estimated, and the three model comparison methods (DIC, CPO, and PPMC) were then used to compare those models. A preferred model was selected based on each method. The effectiveness of these three methods was measured by the number of times the *Mg* was selected as the preferred model across the 20 replications.

In addition to test-level model comparisons within replications, item-level results for both CPO and the discrepancy measures used with the PPMC method were examined. While test-level results are useful in the model comparison process, item-level results may provide diagnostic information regarding the specific misfitting items, which in turn may suggest alternative models.

Bayesian Estimation of Different Models

All the GR models in this study were estimated using MCMC methods and WinBUGS 1.4 (Spiegelhalter et al., 2003). For the standard 2P-GR model, the

following priors were specified: $\theta_i \sim$Normal$(0, 1)$ for Person i, and $a_j \sim$Log normal$(0, 1)$, $b_{j1} \sim$Normal$(0, 0.25)$, and $b_{j(k+1)} \sim$Normal$(0, 0.25)I(b_{jk})$ for Item j, where the notation $I(b_{jk})$ indicates that the threshold $b_{j(k+1)}$ was sampled to be larger than b_{jk} as required under the GR model. For the 1P-GR model, all priors were the same as for the 2P-GR model except that one slope was estimated. For the RS model, the priors were $\theta_i \sim$Normal$(0, 1)$ for Person i, $a_j \sim$Log normal$(0, 1)$ and $b_j \sim$Normal$(0, 0.25)$ for Item j, and $c_1 \sim$Normal$(0, 0.25)$ and $c_{(k+1)} \sim$Normal $(0, 0.25)I(c_k)$ with a constraint of $\sum_k c_k = 0$. It should be noted that, consistent with WinBUGS, precision parameters that are the inverse of variance parameters were used in these prior distribution specifications. These prior specifications were similar to those used in previous research (e.g., Kang et al., 2009).

For the 2D simple-GR model, the prior distributions for item parameters were specified as for the 2P-GR model. The abilities on two dimensions were assigned multivariate normal priors, with means of 0 and variances of 1, and the correlation between two abilities was assigned a normal prior with mean equal to the true correlation of 0.6 and variance of 0.25. This approach was used to address the metric indeterminacy problem in a manner similar to that used by Yao and Boughton (2007).

For estimating the 2D complex-GR model, it is important to solve both metric indeterminacy and rotational indeterminacy. As for 2D simple-GR model, the metric indeterminacy problem was addressed by assigning the abilities on two dimension means of 0 and variances of 1. To solve the rotational indeterminacy issue, the two ability axes were constrained to be orthogonal, and the slope parameters for the nuisance dimension were fixed at 0 for the last 10 items (items measuring the dominant dimension only). Thus, any relationship between the dimensions was derived from the items that were common to both dimensions rather than from a specified correlation between dimensions (see Table 2, Condition 3). Priors for abilities followed a multivariate normal distribution with means 0 and a variance–covariance matrix equal to the identity matrix. These approaches were similar to those used by Bolt and Lall (2003). Prior distributions for other item parameters were the same as defined for the 2P-GR model. It should be noted that although use of an orthogonal solution affects the evaluation of item parameter recovery, model-based expectations (i.e., response category probabilities) are invariant to rotations (e.g., orthogonal vs. oblique). Thus, the solution for rotational indeterminacy should not affect results evaluating model fit and model comparisons.

For the testlet GR model, prior distributions for the item parameters and the examinees' abilities were specified as for the GR models. The testlet effect was assigned a normal prior with mean of 0 and random variance of $\sigma^2_{d(j)}$. As in Bradlow et al. (1999) and Li et al. (2006), the hyperparameter $\sigma^2_{d(j)}$ was specified as an inverse chi-square distribution with degrees of freedom equal to 0.5, indicating a lack of information about this parameter.

Convergence of the Markov chain for each GR model was evaluated using multiple diagnostics, such as history plots, autocorrelation plots, and multiple chains to ensure that target posterior distributions were being sampled (Kim & Bolt, 2007).

Based on the results (see Zhu, 2009, for details), one chain of 5,000 iterations was run for the 2P-GR, 1P-GR, RS, and testlet GR models, and the first 3,000 iterations were discarded. The remaining 2,000 iterations were thinned by taking every other iteration to reduce serial dependencies across iterations. The resulting 1,000 iterations formed the posterior sample and were used in estimating model parameters and conducting model comparisons. For the 2D simple-GR and 2D complex-GR models, a longer chain of 8,000 iterations was run. The first 5,000 iterations were discarded, and the remaining 3,000 iterations were thinned by taking every third iteration to obtain a posterior sample of size 1,000.

Convergence was also assessed by examining parameter recovery for each model using the root mean square error (RMSE) between the true (i.e., generating) and posterior estimates or mean values from the posterior distributions across the 20 replications. Thus, the RMSEs for a specific model were obtained when the data-generating model was the same as the analysis model (i.e., *Mg* = *Ma*). RMSE results indicated that the parameters were recovered well (see Zhu, 2009, for complete results). For example, the average RMSE across all 15 items was 0.07 for both slope and threshold parameters for the 2P-GR model; for the 2D simple-GR model, the RMSE for the interdimensional correlation was 0.016; and for the testlet GR model, the RMSE for the testlet variance was 0.037. The low RMSE values indicated that a posterior sample of 1,000 was adequate for accurate recovery of item parameters for each GR model.

Computation of Model Comparison Criteria

Estimates of the DIC index for different models were directly available from WinBUGS. The computation of the CPO index was implemented by first computing the CPO at the level of an individual item response (CPO_{ij}) in WinBUGS by using,

$$CPO_{ij} = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{f\left(y_{ij} | \eta_t\right)} \right)^{-1}, \tag{11}$$

where y_{ij} is the response of an examinee i on a particular Item j, T is the total number of MCMC draws after the chain has converged, and $f\left(y_{ij} | \eta_t\right)$ is the likelihood of the observed item response y_{ij} based on the sampled parameter values at draw t. An item-level CPO index for a model was obtained in SAS by taking the product of the values of CPO_{ij} across all examinees: $CPO_j = \prod_{i=1}^{N} CPO_{ij}$, where N is the total number of examinees. In addition, a CPO index for the overall test was computed by taking the product of the item-level CPO_j across all items: $CPO = \prod_{j=1}^{I} CPO_j$, where I is the total number of items on the test. The logarithm of the CPO index was used for comparing

models in this study. Recall that the preferred model was the model with the smaller DIC value or the larger CPO value.

The different models in each condition were also compared using the PPMC method. All six discrepancy measures discussed previously were used with PPMC for Condition 1. However, for Conditions 2 to 4, only the two pair-wise discrepancy measures (*global OR* and *Yen's* Q_3 index) were used since these types of measures have been found most useful for detecting multidimensionality or local dependence (e.g., Levy et al., 2009; Sinharay et al., 2006; Zhu & Stone, 2011). To compare different models using PPMC, the frequency of extreme PPP values (i.e., values <0.05 or >0.95 as discussed earlier) for the selected measures was computed for each model across items or item pairs. When *Ma* = *Mg*, it was expected that no or few extreme PPP values would be observed. When the alternative model was estimated, however, more items or item pairs with extreme PPP values would be expected. The preferred model was the model with the fewest number of items and item pairs with extreme PPP values. Note that among all the six discrepancy measures used in this study, three measures (the *item score distribution*, the *global OR*, and *Yen's Q3*) and their PPP values were computed in WinBUGS, and the other three measures (the *item–total score correlation*, *Yen's Q1*, and the *pseudo-count fit statistic*) were computed in SAS. The interested reader can consult (Zhu, 2009) for the SAS and WinBUGS code to compute these various measures.

Results

Condition 1 (2P-GR vs. 1P-GR vs. RS Models)

Table 3 presents the average values (Mean) for the model comparison indices as well as the frequencies (Freq) that each model was selected as the preferred model across the 20 replications. For the item-level discrepancy measures (i.e., *item score distribution*, *item–total score correlation*, *Yen's Q1*, *pseudo-count fit statistic*), the means represent the average numbers of extreme PPP values (i.e., <0.05 or >0.95) across the 15 items and the 20 replications. For the pairwise measures (*global OR* and *Yen's Q3*), the means represent the average number of extreme PPP values across the 105 item pairs and replications. Note that these values were all rounded to the nearest integer.

As shown in this table, the mean DIC value was the smallest for the true model (i.e., 2P-GR) and the largest for the 1P-GR model. Similarly, the mean test-level CPO value was the largest for the true model and the smallest for the 1P-GR model. Thus, both DIC and CPO results indicated that the 2P-GR model fit the data better than the RS model, which in turn fit the data better than the 1P-GR model. Moreover, the generating or true model was consistently selected as the preferred model across the 20 replications using the DIC and CPO indices.

Regarding the PPMC results in Table 3, the 2P-GR model had the lowest average values for all discrepancy measures, providing evidence that the 2P-GR model was preferred. For example, for the *item-total correlation* measure, no item had extreme PPP values for the true model (Mean = 0). However, across the 15 items, there were

Table 3. Model Selection Results Across Items or Item Pairs (Overall Test): Condition 1

	DIC		CPO		Item score distribution		Item–total correlation		Yen's Q_1		Pseudo-count fit statistic		Global OR		Yen's Q_3	
Model	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq
2P-GR[a]	73960	20	−16076	20	0	20	0	20	0	20	0	20	6	20	7	20
1P-GR	75498	0	−16405	0	0	20	12	0	10	0	10	0	79	0	48	0
RS	74484	0	−16190	0	14	0	13	0	8	0	13	0	31	0	13	0

Note. DIC = deviance information criterion; CPO = conditional predictive ordinate; GR = graded response; RS = rating scale model; Freq = frequency; OR = odds ratio; 2P-GR = two-parameter GR model; 1P-GR = one-parameter version of the GR model.

a. The model was the data-generating (i.e., true) model.

an average of 12 items with extreme PPP values for the 1P-GR model, and an average of 13 items with extreme PPP values for the RS model. For *Yen's* Q_3 measure, across 105 item pairs an average of 7 pairs had extreme PPP values for the true model, but an average of 48 and 13 item pairs had extreme PPP values for the 1P-GR and RS models, respectively. In addition, all measures, except the *item score distribution* measure, consistently identified the true model as the preferred model over the 20 replications. Using the *item score distribution* measure, both the 2P-GR and 1P-GR models were identified as preferred models since no items with extreme PPP values were observed for either model. This result indicated that this measure may be ineffective in discriminating between 2P- and 1P-GR models. Since this measure does not compare observed and expected values at different discrete values of ability, as in *Yen's* Q_1 and the *pseudo-count fit statistic*, it is not sensitive to discrepancies between observed and expected values that may exist at specific ability levels. Furthermore, positive and negative discrepancies may cancel each other when the discrepancies are collapsed across ability levels as in the *item score distribution statistic*.

The previous comparisons focused on the selection of a preferred model for the overall test. Although overall model comparisons are useful, comparing models at the item level can provide additional diagnostic information. For example, information regarding item-level fit could be used to identify a subset of items that should be reexamined, modified, or replaced. Among the Bayesian comparison methods considered, CPO and PPMC can be used to compare the models at both test- and item-levels. For Condition 1, the item-level CPO results (not shown here) indicated that the 2P-GR model was preferred for each of the 15 items. With respect to the item-level PPMC results, median PPP values for all four item-level discrepancy measures were around 0.50 for the 2P-GR model. In contrast, extreme PPP values were observed for a majority of items when the 1P-GR or RS models were estimated with all but the *item score distribution* discrepancy measure. The differences in the numbers of extreme PPP values between the models indicated that the 2P-GR was preferred.

It is worthwhile to note that both the DIC and CPO indices are used to compare the relative fit of different models. Thus, they may be used to identify which model among a set of candidate models provides a better fit to the data, but they may not be used to evaluate the degree of fit in an absolute sense. When one or more models among the set of candidate models are appropriate, the better fitting model can be selected by using either DIC or CPO index. However, when the models to be compared are not appropriate or do not fit the data, using these indices may be misleading. For example, if only the 1P-GR and RS models were compared using the DIC or CPO indices for Condition 1, the RS model would be preferred over the 1P-GR model based on the results in Table 3. However, the RS model is not really appropriate since the true model was the 2P-GR model.

In contrast to the DIC and CPO indices, the PPMC method can be also used to evaluate the absolute fit of different models. As shown in Table 3, all effective discrepancy measures had a number of extreme PPP values for the 1P-GR and RS models,

Table 4. Model Selection Results Across Items (Overall Test): Conditions 2 to 4

		DIC		CPO		PPMC (global OR)		PPMC (Yen's Q_3)	
Condition	Model	Mean	Freq	Mean	Freq	Mean	Freq	Mean	Freq
2	2D simple-GR[a]	75434	20	−16430	20	7	20	4	20
	2P-GR	78854	0	−17143	0	75	0	102	0
3	2D complex-GR[a]	71905	20	−15645	20	4	18	4	20
	2P-GR	72093	0	−15670	0	8	2	15	0
4	Testlet GR[a]	74170	20	−16145	20	6	18	5	20
	2P-GR	74924	0	−16287	0	10	2	21	0

Note. DIC = deviance information criterion; CPO = conditional predictive ordinate; PPMC = posterior predictive model checking; OR = odds ratio; GR = graded response; 2P = two parameter; 2D = two-dimensional; Freq = frequency.

a. The model was the data-generating (i.e., true) model for each condition.

indicating that neither of these models fit the simulated 2P-GR data. As expected, no or few PPP values were extreme for the estimated 2P-GR model. This indicated that the 2P model was not only preferred over the other two models, but it also fit the data.

Condition 2 (2P-GR vs. 2D Simple-GR Models)

Table 4 (see Condition 2) presents the results comparing the estimation of a unidimensional 2P-GR model with a 2D simple-GR model (true or generating model) for the simulated 2D simple-structure data. The table provides mean values for each model comparison index, as well as the frequencies the true model (i.e., 2D simple-GR) was preferred across the 20 replications.

The lower DIC value and the higher CPO value for the 2D simple-GR model indicated that the true model was preferred for the overall test. For the PPMC results, on average, only 7 out of 105 item pairs with extreme PPP values for the *global OR* measure (or 4 for *Yen's* Q_3 index) were observed when the true model was used to analyze the data. However, when the unidimensional 2P-GR model was estimated, there were a large number of item pairs with extreme PPP values—75 and 102 pairs for the *global OR* and *Yen's* Q_3 index, respectively. Thus, the PPMC results indicated that the two-dimensional model was preferred over the unidimensional GR model and also indicated that this model provided adequate model fit with respect to associations among item-pair responses. Also, it is apparent that the three methods appeared to perform equally well with regard to the frequency of selecting the two-dimensional GR model as the preferred model at the overall test level. All the indices identified the true model for each of the 20 replications.

As for Condition 1, item-level results were examined to diagnose any particular source of model misfit. The item-level CPO results are presented in Table 5 (see

Table 5. Item-Level CPO Index Results for Each Item: Conditions 2 to 4

	Condition 2			Condition 3			Condition 4		
Item	2D Simple-GR[a]	2P-GR	Freq	2D Complex-GR[a]	2P-GR	Freq	Testlet GR[a]	2P-GR	Freq
1	−1274	−1291	20	−1201	−1205	19	−1269	−1269	15
2	−1162	−1209	20	−1238	−1242	19	−1293	−1294	15
3	−998	−1077	20	−1209	−1212	19	−1266	−1267	16
4	−1214	−1231	20	−1076	−1081	20	−1119	−1120	14
5	−1001	−1044	20	−1071	−1076	20	−1204	−1204	16
6	−999	−1075	20	−1114	−1115	14	−1121	−1162	20
7	−1301	−1319	20	−1140	−1139	10	−1151	−1193	20
8	−1133	−1180	20	−1113	−1114	13	−1124	−1166	20
9	−848	−923	20	−983	−984	17	−986	−987	18
10	−1209	−1227	20	−979	−978	15	−986	−987	16
11	−1133	−1184	20	−961	−959	14	−968	−970	18
12	−1032	−1114	20	−988	−989	14	−991	−993	16
13	−1277	−1297	20	−958	−959	14	−962	−965	19
14	−1002	−1049	20	−807	−808	18	−811	−814	20
15	−845	−923	20	−806	−808	12	−814	−816	20

Note. CPO = conditional predictive ordinate; GR = graded response; 2P = two-parameter; 2D = two-dimensional; Freq = frequency.

a. The model was the data-generating (i.e., true) model for each condition.

Condition 2), including the mean CPO values for each estimated model as well as the frequencies the true model was selected as the preferred model across the 20 replications. For each item, the average CPO value (across the 20 replications) for the two-dimensional model was larger than the value for the unidimensional model. This indicated that the two-dimensional model, which was preferred for the overall test, was also preferred for each item. The large differences for the item-level CPO values between the two models provided strong evidence that the two-dimensional model fit each item significantly better than the unidimensional model. In addition, the two-dimensional model was chosen as the preferred model for each of the 20 replications.

With regard to the PPMC results, pie plots were used rather than tables of PPP values to examine and explore patterns in item-level results. As an example, the pie plots in Figure 1 (Condition 2) provide median PPP values for *Yen's* Q_3 measure when items responses were simulated under a 2D simple-GR model and two competing models were estimated: the 2P-GR model and the 2D simple-GR model. In each plot, there is one pie for each item pair, and the proportion of a circle that is filled corresponds to the magnitude of median PPP value across the 20 replications. An empty pie reflects a PPP-value of 0.0, whereas a fully filled pie reflects a PPP value of 1.0. As shown from the pie plots, when a unidimensional GR model was estimated (upper left plot), median PPP values for all item pairs were extreme and the items fell into two clusters—Items 1 to 8 in one, and Items 9 to 15 in another. This pattern

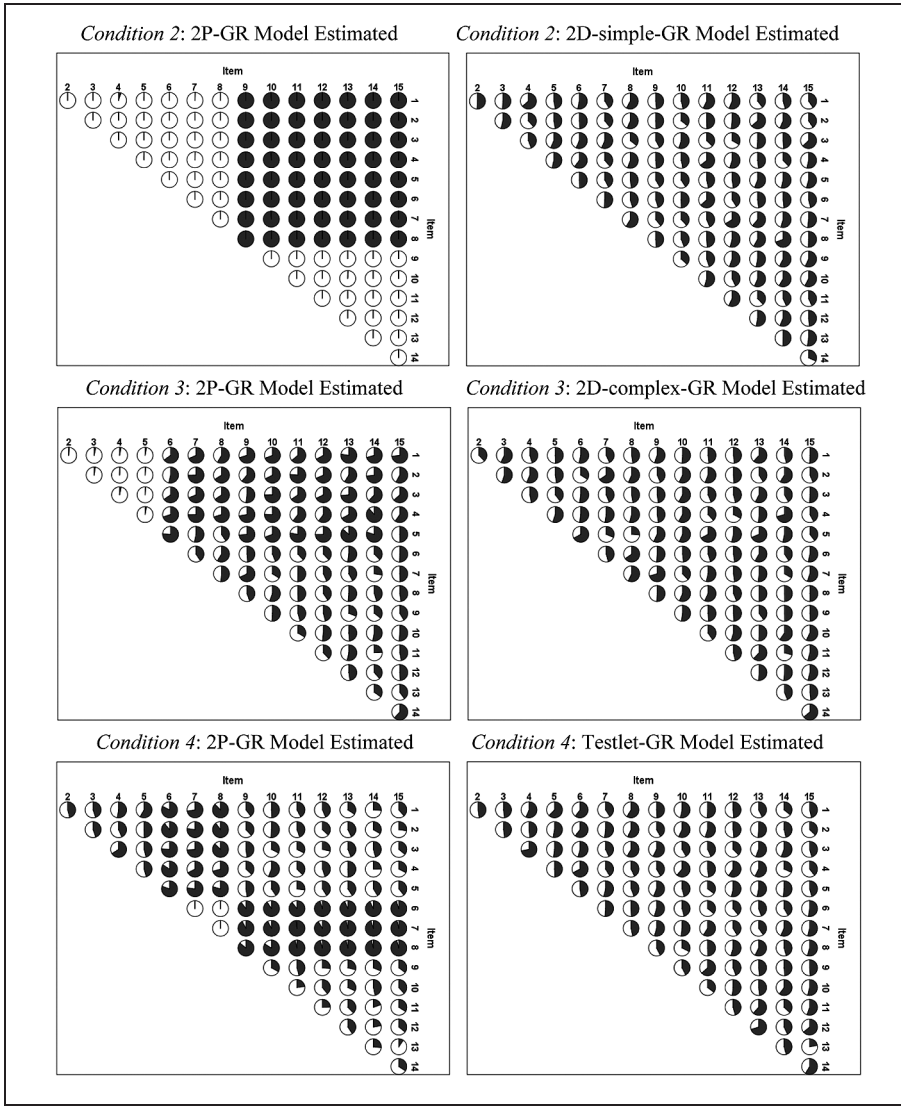

Figure 1. Display of median PPP values for item pairs for *Yen's* Q_3 for Conditions 2 to 4
Note. PPP = posterior predictive *p*; 2P = two-parameter; GR = graded response; 2D = two-dimensional.

indicated that a unidimensional model did not fit the data and a two-dimensional model might be considered more appropriate to model the data. As expected, when a two-dimensional model (i.e., true model) was estimated (upper right plot), median PPP values were around 0.5 for all pairs, suggesting this model fit the data. Pie plots based on the *global OR* measure exhibited a similar pattern.

Condition 3 (2P-GR vs. 2D Complex-GR Models)

In Condition 3, the generated data were 2D with complex structure. Items 1 to 5 measured a dominant dimension as well as a nuisance dimension, and Items 6 to 15 only measured the dominant dimension (see Table 2). A unidimensional 2P-GR model and a two-dimensional GR model were estimated and compared.

The model comparison results for the overall test are presented in Table 4 (see Condition 3). The lower DIC value and higher CPO value for the 2D complex-GR model (true model) indicated that this multidimensional model was preferred over the unidimensional GR model. For the PPMC indices, 4 out of 105 item pairs demonstrated extreme PPP values for both pair-wise measures when the two-dimensional model was estimated. In contrast, when the unidimensional GR model was estimated, more item pairs demonstrated extreme PPP values—8 and 15 pairs for the *global OR* measure and *Yen's* Q_3 index, respectively.

The DIC index, the CPO index, and the PPMC method using *Yen's* Q_3 measure appeared to perform equally well with regard to the frequency of selecting the two-dimensional GR model as the preferred model. Based on the model comparisons for these indices, the true model was selected as the preferred model every time (20) over the 20 replications. However, when the *global OR* measure was used, the unidimensional GR model was selected incorrectly as the preferred model for 2 of the 20 replications. This example again illustrates that the choice of the discrepancy measure may affect the performance of the PPMC application in comparing different models.

The item-level CPO results (means and frequencies) for this condition are presented in Table 5 (see Condition 3). For Items 1 to 5, which measured both the dominant and nuisance dimensions, the item-level CPO identified the generating two-dimensional model as the preferred model for 19 to 20 times over the 20 replications. However, for the items that only measured the dominant dimension (Items 6-15), the two-dimensional model was identified as the preferred model with lower frequencies (10 to 18 over the 20 replications). This would be expected since the unidimensional GR model should be appropriate for these items.

It may be worthy to note that the differences between the item-level CPO values were quite small for the two models, particularly for the items reflecting one dimension (Items 6-15). Spiegelhalter et al. (2003) discussed that any difference in DIC less than 5 units for two models may not indicate sufficient evidence to favor one model over another. Though there are no discussed guidelines available for CPO, the item-level CPO results for this condition may indicate that a difference of less than 3 units may not provide sufficient evidence supporting one model over another. However, the amount of difference in CPO necessary to suggest a significant difference between models needs further investigation.

The pie plots in Figure 1 (Condition 3) display median PPP values for *Yen's* Q_3 measure for this condition. All the PPP values were around 0.5 when the true 2D complex-GR model was estimated (middle right plot) and therefore provided evidence of model fit. In contrast, when the unidimensional GR model was estimated

(middle left plot), all the PPP values were extreme for the item pairs involving the first five items, but around 0.5 for the other item pairs. This pattern indicated that the unidimensional GR model was not appropriate for Items 1 to 5, but was appropriate for Items 6 to 15. Additionally, the close to 0 PPP values for the item pairs among Items 1 to 5 indicated that the realized correlations among these five items were consistently larger than the predicted correlations under the unidimensional GR model.

Condition 4 (2P-GR vs. Testlet GR Models)

In Condition 4, the responses to a test with locally dependent items (testlet) were simulated. Items 6, 7, and 8 were designed to reflect a testlet with moderate dependence among the items. Table 4 (see Condition 4) presents the results for comparing a GR model for testlets with a unidimensional GR model. The lower DIC value, the higher CPO value, and the fewer item pairs with extreme PPP values (i.e., values <0.05 or >0.95) for the testlet GR model all indicated that this complex model fit the overall test better than the unidimensional GR model. Furthermore, the DIC index, the CPO index, and the PPMC method using *Yen's* Q_3 measure appeared to perform equally well. Based on these measures, the testlet GR model was selected as the preferred model every time (20) over the 20 replications. However, when the *global OR* measure was used with PPMC, the testlet GR model was chosen as the preferred model for 18 of the 20 replications. As for Condition 3, use of *Yen's* Q_3 as a discrepancy measure appeared to be slightly more effective than use of the *global OR* measure.

Table 5 (Condition 4) presents results for the item-level model comparisons. For the items in the testlet (Items 6, 7, and 8), the mean CPO values for the testlet GR model were much larger (~42 units) than the values for the unidimensional model, and the testlet GR model was chosen as the preferred model for all the 20 replications. For the simulated locally independent items, the mean CPO values for the two models were very close. The mean CPO values for the testlet GR model were less than 3 units larger than the values for the unidimensional GR model. Based on the CPO results for Condition 3, less than 3 units in difference may indicate insufficient evidence for selecting the complex testlet GR model as the preferred model. Nonetheless, if these small differences were considered evidence in favor of the testlet model, the testlet GR model would be selected as the preferred model for these items for 14 to 20 replications. Thus, more research is needed to determine what differences in CPO values should be used to select one model over another.

The two pie plots at the bottom of Figure 1 (see Condition 4) illustrate the median PPP values for *Yen's* Q_3 measure for the testlet and unidimensional GR models, respectively. As can be observed, when the testlet GR model was estimated (bottom right plot), all PPP values were around 0.5, suggesting model-data-fit for the testlet GR model. In contrast, when the unidimensional GR model was estimated (bottom left plot), all PPP values were extreme for item pairs reflecting local dependence (Items 6, 7, and 8), but around 0.5 for pairs among simulated locally independent

items. Additionally, PPP values of nearly 0 for item pairs comprising the testlet indicated that the realized correlations among these items were consistently larger than predicted correlations under the unidimensional GR model. These results indicated that the unidimensional GR model was not appropriate for Items 6, 7, and 8, but was appropriate for the other items.

It should be noted that results for the 2D complex-structure model (Condition 3) and the testlet model (Condition 4) were very similar. As noted by a reviewer, this was because of the similarity between the 2D complex-structure model used herein and the testlet model. In both models all items measure a single primary dimension while a subset of items in both the 2D complex-structure and testlets models are related further because of another source of shared variance.

Discussion

The purpose of this study was to evaluate the relative performance of three Bayesian model comparison methods (DIC, CPO, and PPMC) in selecting a preferred model among a set of alternative GR models for performance assessment applications. The alternative models considered in this study reflected sources of potential misfit and reflected more complex IRT models that might be theoretically more appropriate for some types of performance assessment data but require Bayesian estimation methods. For the conditions examined in this study, the results of this study indicated that these three methods appeared to be equally accurate in selecting the true model as the preferred model when considering all items simultaneously (test level). However, CPO and PPMC were found to be more informative than DIC since information about model fit was also available at the item level.

Consistent with previous studies (Li et al., 2006; Sinharay, 2005), the PPMC approach was found to be effective for performing model comparisons in the performance assessment context. Moreover, an advantage of PPMC applications is that they can be used to compare both the relative and absolute fit of different models. In contrast, the DIC and CPO model comparison indices only consider the relative fit of different models. If the true model is among the candidate models, DIC or CPO can be used to select the preferred model that should also fit the data. However, in practice, it is not known whether the true model is included in the set of candidate models. Thus, it is important that the assessment of absolute model fit be combined with model comparison methods. Ideally, model selection (e.g., DIC/CPO) and model fit (PPMC) methods should both be used for real testing applications. As a result of their simplicity, DIC or CPO could be first used to choose a preferred model from a set of candidate models. Using multiple discrepancy measures, the PPMC method could then be applied to evaluate the fit of the selected model at the test and item levels. In this sense, these methods are not competing in nature, but are instead complementary.

Contrary to previous studies, DIC, CPO, and PPMC appeared to perform equally well in this study, whereas their performance has been found in previous research to depend on specific conditions as well as models compared. This finding may be due

to the large sample size used in this study that helped ensure accuracy in model parameter estimation. As discussed in Kang et al. (2009), sample size may affect the behavior of model comparison indices. For example, the performances of AIC and BIC in their study were also quite similar with a large sample size. They also found that a large sample size was required to select the correct model for DIC and PsBF.

For PPMC applications, results from this study also indicate that the choice of discrepancy measures affects the performance of model comparison methods. If the specific discrepancy measure is not effective, PPMC will be less effective than DIC and CPO. As shown in Conditions 3 and 4, when *Yen's* Q_3 measure was used with PPMC, the PPMC index performed equally well with DIC and CPO. However, the performance of the *global OR* measure with PPMC was less effective than using the DIC and CPO indices.

It is worthy to note that the results from this study also offer implications for using test-level versus item-level model comparisons in testing applications. In Condition 4, for example, the test-level model comparisons favored the testlet GR model whereas the item-level model comparisons favored the testlet GR model only for a subset of items. Mixed item formats are often included in testing applications (e.g., multiple-choice items and constructed response items), and in these types of applications, different IRT models may be estimated for different subsets of items. Similarly in performance assessments, different models could be estimated for different subsets of items based on theoretical or substantive grounds. A subset of items in a performance assessment may use a common stimulus, which may cause local dependence among the items whereas no common stimuli may be used in the remaining items. Thus, a mixed modeling approach using both a testlet GR model and a GR model could be employed in an IRT application to item responses.

Although comparing models for sets of items (test level) is often the focus, item-level results may provide added value to the analysis. For example, the test-level CPO index is a summary index across all items. If most of the individual items reflect small differences on item-level CPO values between two competing models, the aggregated test-level CPO difference might be misinterpreted. For this case, if the focus is on the test-level CPO results, a more complex model may be favored over a simpler model. But if item-level CPO results are further examined, small CPO difference for each individual item may suggest that the simpler model is adequate and model parsimony may favor the use of the simpler model for the overall test.

While item-level results may provide useful diagnostic information regarding the source of misfit, the specific misfitting items, and possible alternative models, the use of item-level results complicates the discussion by the possibility that different models may be suggested for different items. For example, the PPMC results for Condition 3 indicated that the misfit of a unidimensional GR model to the responses to the first five items was due to the higher than expected correlations among these items. This might suggest a possible alternative model in which a second dimension is considered for these five items. However, it is also important that substantive knowledge about the items and relationships between items be considered when

selecting a final model. For example, if a testlet GR model is indicated for a subset of items, and there is no substantive basis for the observed local dependence, the subset of items could be examined to identify the source of apparent local dependence. Similarly, test items could be examined for construct-irrelevant variance given any observed multidimensionality in the item responses.

Though the conditions in this study were carefully designed, some factors were fixed at values realistic to typical performance assessments. Therefore, results may not generalize to other conditions not considered herein. For example, this study was limited in terms of the length of tests (15 items), number of response categories (5 categories), specific polytomous model estimated (graded model), and number of modeled dimensions (2 dimensions). Future research could consider different sample sizes or test lengths to evaluate the effect of sample size and test length on the performance of these model comparison methods. Other factors, such as the number of dimensions, other multidimensional structures, and varying the testlet effect could be also explored. Higher correlations between dimensions could be considered as correlations may be higher in some performance assessment applications. For the PPMC method, a *global OR* was used in this study. Since the dichotomization used in the *global OR* may likely result in loss of information, other more informative OR measures could be employed for model comparison. For example, a conditional OR (Mantel–Haenszel) statistic (Agresti, 2002), or the Liu–Agresti cumulative common OR (Liu & Agresti, 1996) for ordinal variables, could prove more powerful than the *global OR* for evaluating the unidimensionality or local independence assumptions for polytomous items.

Another possible limitation to this study was that 20 replications at each combination of experimental conditions were used to compare their performance. Although this was a small number, it was consistent with previous research focusing on Bayesian methods. Furthermore standard deviations across replications indicated relatively stable results with low variability in indices across replications. For example, the mean and standard deviation for the CPO index across 20 replications for simulated two-dimensional item responses were −16430.11 and 57.02, respectively. Although more replications could be used to obtain even more reliable results, the general patterns in the results that were observed would not likely change with additional replications. Last, although the design of the study reflects a performance assessment context, model comparison results should extend to other testing applications that use polytomously scored items (e.g., psychological assessments). For example, model comparisons involving the RS model or multidimensional models have direct implications for these types of applications. Also, model comparisons in this study revolved around item responses based on more complex models since the constructs measured in real performance assessments usually are complex. Future research could consider conditions in which the generating model is a less complex model (e.g., RS model) than the comparison model (e.g., 2P graded model). Such comparisons would provide additional information about the use of the model comparison methods in other testing applications, such as psychological assessments with RSs.

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

Agresti, A. (2002). *Categorical data analysis*. Hoboken, NJ: John Wiley.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

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

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395-414.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Chen, W., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

De Ayala, R. J. (1994). The influence of dimensionality on the graded response model. *Applied Psychological Measurement*, *18*, 155-170.

Fu, J., Bolt, D. M., & Li, Y. (2005). *Evaluating item fit for a polytomous fusion model using posterior predictive checks*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153-160.

Gelfand, A. E., Dey, D. K., & Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 147-167). Oxford, England: Oxford University Press.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* New York, NY: Chapman & Hall.

Kang T., & Cohen, A. S. (2007). IRT model selection methods for dichotomous items. *Applied Psychological Measurement*, *31*, 331-358.

Kang T., Cohen, A. S., & Sung, H-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499-518.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795.

Kim, J., & Bolt, D. (2007). Estimating item response theory models using Markov Chain Monte Carlo methods. *Educational Measurement: Issues and Practice, 26*(4), 38-51.

Lane, S., & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 387-424). Westport, CT: American Council on Education/ Praeger.

Lane, S., Stone, C. A., Ankenmann, R. D., & Liu, M. (1995). Examination of the assumptions and properties of the graded item response model: An example using a mathematics performance assessment. *Applied Measurement in Education, 8*, 313-340.

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, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement, 30*, 3-21.

Liu, I.-M., & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometricis, 52,* 1223-1234.

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement, 14*, 59-71.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics, 12*, 1151-1172.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement, 17*.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42*, 375-394.

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology, 59*, 429-449.

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

Spiegelhalter, D. J., Best, N., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, 64*, 583-640.

Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. (2003). *WINBUGS Version 1.4 user's manual* [Computer software manual]. Cambridge, England: MRC Biostatistics Unit.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement, 37*, 58-75.

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement, 26*, 109-128.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83-105.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.

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

Zhu, X. (2009). *Checking fit of item response models for performance assessments using Bayesian analysis*. Unpublished dissertation. University of Pittsburgh.

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