Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Posterior Predictive Model Checking for Conjunctive Multidimensionality in Item Response Theory

Roy Levy

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS published online 19 August 2011
DOI: 10.3102/1076998611410213

The online version of this article can be found at:
http://jeb.sagepub.com/content/early/2011/08/19/1076998611410213

A more recent version of this article was published on - Oct 14, 2011

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Version of Record - Oct 14, 2011

>> Proof - Aug 19, 2011

Journal of Educational and Behavioral Statistics
Month 2011, Vol. 000, No. 00, pp. 1–23
DOI: 10.3102/1076998611410213

Posterior Predictive Model Checking for Conjunctive Multidimensionality in Item Response Theory

Roy Levy
Arizona State University, Tempe, AZ, USA

If data exhibit multidimensionality, key conditional independence assumptions of unidimensional models do not hold. The current work pursues posterior predictive model checking (PPMC) as a tool for criticizing models due to unaccounted for dimensions in data structures that follow conjunctive multidimensional models. These pursuits are couched in previous work investigating factors influencing dimensionality and dimensionality assessment. A simulation study investigates the model checking tools in the context of item response theory (IRT) for dichotomous observables, in which a unidimensional model is fit to data that follow a conjunctive multidimensional model. Key findings include (a) support for the hypothesized effects of the manipulated factors and (b) the superiority of certain discrepancy measures for conducting PPMC for dimensionality assessment.

Keywords: *posterior predictive model checking; dimensionality assessment; item response theory; multidimensionality; local independence*

Although Bayesian modeling and estimation strategies are receiving a considerable amount of attention in item response theory (IRT; e.g., Bradlow, Wainer, & Wang, 1999; Jannsen, Tuerlinckx, Meulders, & De Boeck, 2000; Patz & Junker, 1999), model diagnostics and model criticism remain relatively unexplored aspects of Bayesian psychometric modeling. This study investigates the use of posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Meng, 1994; Rubin, 1984) to criticize unidimensional models when the data follow a conjunctive multidimensional structure. PPMC has seen applications and has been the subject of methodological investigations in IRT (e.g., Hoijtink, 2001; Jannsen et al., 2000; Sinharay, 2005, 2006; Sinharay, Johnson, & Stern, 2006). Central to the current work, Levy, Mislevy, and Sinharay (2009) investigated the use of PPMC to critique unidimensional models when the data follow compensatory multidimensional structures. More specifically, Levy et al. (2009) fit the unidimensional two parameter logistic model (2-PLM) for dichotomous observables (i.e., scored item responses):

$$P(X_{ij} = 1|\theta_i, b_j, a_j) = \frac{\exp(a_j(\theta_i - b_j))}{1 + \exp(a_j(\theta_i - b_j))}, \quad (1)$$

where X_{ij} is the observable response (coded as 0 or 1) from examinee i to item j, θ_i is the latent variable for examinee i, and b_j and a_j are difficulty and discrimination parameters, respectively, for item j. The current work considers the case of fitting the 2-PLM to data that follow a conjunctive multidimensional model, given by

$$P(X_{ij} = 1|\boldsymbol{\theta}_i, \mathbf{b}_j) = \prod_{m=1}^{M} \frac{\exp(\theta_{im} - b_{jm})}{1 + \exp(\theta_{im} - b_{jm})}, \quad (2)$$

where $\theta_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iM})'$ is a vector of M latent variables that characterize examinee i and $\mathbf{b}_j = (b_{j1}, b_{j2}, \ldots, b_{jM})'$ $(\theta_i = \theta_{i1}\theta_{i2}, K, \theta_{iM})'$ is a vector of M parameters for item j corresponding to the difficulties along the M dimensions (e.g., Embretson, 1997). Conjunctive multidimensional IRT (MIRT) models have considerable appeal in terms of their connections to cognitive underpinnings of how examinees approach and solve tasks (Embretson, 1984, 1997). Hence, we pursue the situation in which a conjunctive MIRT model characterizes the response process but our model assumes unidimensionality. We consider multidimensional structures in which (a) all the observables reflect a primary dimension but (b) some observables additionally reflect one of several auxiliary dimensions. A variety of phenomena, including differential item functioning, item drift, testlet effects, method effects, and the influence of multiple aspects of proficiency may be framed in terms of multidimensionality and, if unmodeled, result in violations of assumed local independence structures (e.g., Bolt & Stout, 1996; Bradlow et al., 1999; Stout et al., 1996; Yen, 1993).

Factors Influencing Dimensionality Assessment

Levy et al. (2009) considered the case of fitting the 2-PLM to data that follow compensatory MIRT models. As in the present study, they considered structures in which there is one primary dimension that influences all items and additional auxiliary dimensions that influence some of the items. Their principal findings concern several key factors that influence the ability to detect the presence of multidimensionality:

- As the strength of the dependence of items on the auxiliary dimensions increases, it becomes easier to detect data-model misfit in terms of pairs of items that reflect the same multiple dimensions.
- As the correlations among the dimensions increases, it becomes harder to detect data-model misfit in terms of all types of item pairs.
- As the proportion of items reflecting an auxiliary dimension increases, it becomes harder to detect data-model misfit in terms of items that reflect the same multiple

dimensions but easier to detect data-model misfit in terms of item pairs that reflect different multiple dimensions and, to a lesser extent, item pairs that reflect the primary dimension only.

Levy et al. (2009) motivated their investigations and interpreted their findings in terms of conditional covariance theory for compensatory and generalized compensatory models (Stout et al., 1996; Zhang & Stout, 1999). However, the conjunctive MIRT model is not a member of the class of generalized compensatory models, and no comparable foundation has been established for conjunctive models. Hence, it is an open question whether the implications of theoretical work on generalized compensatory models holds for conjunctive MIRT models.

The primary objective of this article is to pursue the utility of PPMC for IRT models when data structures exhibit conjunctive multidimensionality. We investigate whether the factors that affect dimensionality assessment in compensatory data structures similarly influence dimensionality assessment in the current context and we compare the performance of several alternative discrepancy measures that have been proposed for the assessment of local dependence and multidimensionality. To these ends the remainder of this article is organized as follows. The next section describes PPMC for performing model criticism. The remaining sections describe and discuss the results of a simulation study investigating the efficacy of PPMC for criticizing unidimensional models in light of conjunctive multidimensionality.

PPMC

Let $\boldsymbol{\theta}$, $\boldsymbol{\omega}$, and $\mathbf{X}$ denote the complete collections of subject variables, item parameters, and observables, respectively. The complete collection of unknown model parameters is given by $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\omega})$. The posterior distribution for $\boldsymbol{\Omega}$ given $\mathbf{X}$ is

$$P(\boldsymbol{\Omega}|\mathbf{X}) = \frac{P(\boldsymbol{\Omega}) \times P(\mathbf{X}|\boldsymbol{\Omega})}{\int_{\boldsymbol{\Omega}} P(\boldsymbol{\Omega}) \times P(\mathbf{X}|\boldsymbol{\Omega}) d\boldsymbol{\Omega}} \propto P(\boldsymbol{\Omega}) \times P(\mathbf{X}|\boldsymbol{\Omega}). \tag{3}$$

The posterior predictive distribution is the posterior distribution of potential but unobserved data values, denoted $\mathbf{X}^{\text{rep}}$, given by

$$P(\mathbf{X}^{\text{rep}}|\mathbf{X}) = \int_{\boldsymbol{\Omega}} P(\mathbf{X}^{\text{rep}}|\mathbf{X}, \boldsymbol{\Omega}) P(\boldsymbol{\Omega}|\mathbf{X}) d\boldsymbol{\Omega} = \int_{\boldsymbol{\Omega}} P(\mathbf{X}^{\text{rep}}|\boldsymbol{\Omega}) P(\boldsymbol{\Omega}|\mathbf{X}) d\boldsymbol{\Omega}.$$

PPMC (Gelman et al., 1996; Meng, 1994; Rubin, 1984) analyzes characteristics of the observed data and/or the discrepancy between the observed data and model relative to the posterior predictive distribution. Discrepancy measures, each generically denoted by $D(\mathbf{X}, \boldsymbol{\Omega})$, are defined to capture relevant features of the data

and/or the discrepancy between data and the model. Meaningful differences between the realized discrepancies $D(\mathbf{X}, \mathbf{\Omega})$, based on the observed data, and the distribution of $D(\mathbf{X}^{\text{rep}}, \mathbf{\Omega})$, based on the posterior predictive distribution, are indicative of data-model misfit.

The results of PPMC may be summarized by the posterior predictive p value (PPP value; Gelman et al., 1996; Meng, 1994), the tail area of the posterior predictive distribution of the discrepancy measure corresponding to the observed value for the discrepancy measure, given by

$$\begin{aligned}\text{PPP value} &= P(D(\mathbf{X}^{\text{rep}}, \mathbf{\Omega}) \geq D(\mathbf{X}, \mathbf{\Omega})|\mathbf{X}) \\ &= \int I[D(\mathbf{X}^{\text{rep}}, \mathbf{\Omega}) \geq D(\mathbf{X}, \mathbf{\Omega})]P(\mathbf{X}^{\text{rep}}|\mathbf{\Omega})P(\mathbf{\Omega}|\mathbf{X})d\mathbf{X}^{\text{rep}}d\mathbf{\Omega},\end{aligned}$$

where $I[\cdot]$ is the indicator function that takes on a value of one when its argument is true and zero otherwise. Extreme PPP values are indicative of data-model misfit; values near zero (or unity) imply the model is underpredicting (overpredicting) the features captured by the discrepancy measure. In practice, Markov chain Monte Carlo techniques for model estimation may be easily extended to conduct PPMC. Simulated draws from $P(\mathbf{\Omega}|\mathbf{X})$ are employed to generate the $\mathbf{X}^{\text{rep}}$, which are then used in computing the $D(\mathbf{X}^{\text{rep}}, \mathbf{\Omega})$.

PPMC is a powerful and flexible tool for model criticism and has many advantageous properties (Levy et al., 2009). By constructing the reference distribution empirically, PPMC supports model criticism in situations that pose problems for traditional techniques, such as when (a) sample sizes are too small to warrant the use of asymptotic sampling distributions, (b) other regularity conditions associated with frequentist model checking do not hold (e.g., cell frequencies are not large enough to justify assumed distributions; Fu, Bolt, & Li, 2005, Sinharay, 2006), or (c) sampling distributions for the discrepancy measure of interest cannot be determined (Jannsen et al., 2000) or are not well defined (see Chen & Thissen, 1997, for examples in the context of assessing local dependence for IRT models). Furthermore, PPMC incorporates uncertainty of model parameter estimates into the model checking procedure using the full posterior distribution rather than point estimates of the model parameters (Meng, 1994).

Limitations of PPMC include its computational demands and behavior under null conditions. Robins, van der Vaart, and Ventura (2000) showed that PPP values need not be uniformly distributed under null conditions, even asymptotically. The distribution is centered at .5 but may be less dispersed than a uniform distribution (Meng, 1994; Robins et al., 2000). In a hypothesis testing framework, a (two-tailed) test with significance level α may be performed by rejecting the null hypothesis of data-model fit if the PPP value is less than $\alpha/2$ or is greater than $1 - \alpha/2$. The behavior of the PPP values under null conditions implies that such a test could be lead to empirical Type I error rates less than α.

In evaluating IRT models, Sinharay et al. (2006) found PPP values under null conditions to be less dispersed than uniform. Similar results were found by Fu et al. (2005) in a related context. Specifically relevant to the current work, Levy et al. (2009) found considerable differences in the behavior of PPP values for different discrepancy measures under null conditions of unidimensionality. Although several discrepancy measures yielded distributions of PPP values that were underdispersed, the model-based covariance (MBC; Reckase, 1997), Q_3 (Yen, 1993), and Mantel-Haenszel (Sinharay et al., 2006) statistics were quite close to uniformly distributed (Levy et al., 2009).

The current work treats PPMC and PPP values as pieces of statistical evidence for, rather than a test of, data-model (mis)fit (Berkhof, van Mechelen, & Gelman, 2004; Stern, 2000). This is motivated by recommendations on the general practice of using statistical information regarding fit in a larger, theory-guided approach to model criticism (Sinharay, 2005). This orientation is particularly appropriate for the current context of investigating local dependence, as "Any meaningful interpretation of the LD [local dependence] indexes requires skill and experience in IRT analysis and close examination of the item content" (Chen & Thissen, 1997, p. 288) and "the process of interpreting dependence itself is a somewhat imprecise exercise. LID [local item dependence] analyses are largely exploratory in nature, and are completed to provide guidance for the test developer" (Zenisky, Hambleton, & Sireci, 2003, pp. 17–18).

From this perspective, PPMC and PPP values are viewed as diagnostic measures aimed at assessing model strengths and weaknesses rather than whether or not the model is true (Fu et al., 2005; Gelman et al., 1996; Levy et al., 2009). To this end, one advantage of PPMC is that any function of interest may be investigated. Functions should be chosen to reflect the (possibly multiple) features of interest; concluding that a model adequately captures some but not all features of the data is not uncommon. For example, in the context of applying PPMC to detect compensatory multidimensionality, Levy et al. (2009) concluded that the 2-PLM was sufficient to recover the marginal difficulty of the items, but not their associations with one another.

Empirical Study of PPMC for Conjunctive Multidimensionality

The goal of this work is to investigate the effectiveness of different discrepancy measures when PPMC is used to detect the misfit of unidimensional models fit to data exhibiting conjunctive multidimensionality. A Monte Carlo study is conducted in which multiple data sets are generated from three-dimensional conjunctive MIRT models and fit with the 2-PLM to facilitate an examination of the utility of PPMC for diagnosing data-model misfit. All data sets consisted of $J = 32$ items so as to easily manipulate the proportion of items reflecting multiple dimensions. The three latent dimensions consist of the primary dimension, θ_1, that influences all the items and two auxiliary

TABLE 1
Patterns of Multidimensionality

		Number of Items Reflecting Multiple Dimensions					
		4		16		28	
Item	b_{j1}	θ_2	θ_3	θ_2	θ_3	θ_2	θ_3
1	−2.500						
2	−2.000				X		X
3	−1.750			X		X	
4	−1.500						X
5	−1.340					X	
6	−1.170		X		X		X
7	−1.000			X		X	
8	−0.875						X
9	−0.750					X	
10	−0.625				X		X
11	−0.500	X		X		X	
12	−0.400						X
13	−0.300					X	
14	−0.200				X		X
15	−0.100			X		X	
16	0.000						
17	0.000						
18	0.100			X		X	
19	0.200				X		X
20	0.300					X	
21	0.400						X
22	0.500	X		X		X	
23	0.625				X		X
24	0.750					X	
25	0.875						X
26	1.000			X		X	
27	1.170		X		X		X
28	1.340					X	
29	1.500						X
30	1.750			X		X	
31	2.000				X		X
32	2.500						

dimensions, θ_2 and θ_3, which influence a subset of the items. The latent variables are jointly distributed as random samples from a $N(\mathbf{0}, \mathbf{\Sigma})$ population, where the elements along the main diagonal of $\mathbf{\Sigma}$ are all unity and the off-diagonal elements are the correlations. In any condition, the three bivariate correlations were equal;

the chosen values for the correlations span the range from no association (0) to weak (.3) to strong (.7) to extreme (.9).

Table 1 provides a layout of conditions corresponding to the proportion of multidimensional items. The second column gives the value of the location parameter along θ_1 in the data-generating MIRT model. The proportion of items reflecting multiple dimensions were varied from low (4 items) to medium (16 items) to high (28 items). Table 1 indicates which items reflect the additional dimensions in these conditions; the mark "X" indicates that the item reflects the second or third dimension.

To operationalize the notion of strength of dependence on auxiliary dimensions in conjunctive MIRT models, we employ the location parameter along the auxiliary dimension (b_{j2} or b_{j3}) as an indicator of its relevance to the item. Holding all else constant, as the difficulty parameter for an item along a particular dimension decreases, the less influential the dimension is on the item. The values for b_{j2} and b_{j3} (constant over items that depend on θ_2 or θ_3 in any one condition) were varied from -1.0 to -0.5 to 0.5 to 1.0, representing an increasing strength of dependence (for these items) on the auxiliary dimensions. Finally, three sample sizes were investigated: 250, 750, and 2,500. Null conditions of unidimensionality were reported and discussed by Levy et al. (2009) and are integrated into the discussion of the results of this study.

There are 144 combinations of the manipulated factors. For each condition in the study, 50 replications of the following procedures were conducted. Data were generated according to the model specified by the condition, and the 2-PLM model was estimated using standard normal prior distributions for each θ_i, b_j, and $\ln(a_j)$ via a Metropolis-within-Gibbs sampler (Patz & Junker, 1999) to estimate the posterior and posterior predictive distributions. For each analysis, five chains were run from overdispersed starting points for 200 iterations after a burn-in phase. These iterations were thinned by two and the remaining iterations were pooled to yield 500 iterations for use in conducting PPMC.

Discrepancy Measures

A number of the discrepancy measures involve observed and expected frequencies of the bivariate response patterns for a pair of items. The expected frequencies are implied by the IRT model, which may be obtained by integration over the distribution of θ (Chen & Thissen, 1997), which in the current context is the posterior distribution.

Let $n_{kk'}$ denote the number of examinees who have a value of k for item X_j and a value of k' for item $X_{j'}$. X^2 and G^2 discrepancy measures for item pairs (see, e.g., Chen & Thissen, 1997) are given by

$$X^2_{jj'} = \sum_{k=0}^{1} \sum_{k'=0}^{1} \frac{(n_{kk'} - E(n_{kk'}))^2}{E(n_{kk'})}$$

and

$$G^2_{jj'} = -2\sum_{k=0}^{1}\sum_{k'=0}^{1} n_{kk'} \ln \frac{E(n_{kk'})}{n_{kk'}},$$

respectively. Several correlational measures are explored, including the covariance,

$$\mathrm{COV}_{jj'} = \frac{\sum_{i=1}^{N}(X_{ij} - \bar{X}_j)(X_{ij'} - \bar{X}_{j'})}{N} = \frac{(n_{11})(n_{00}) - (n_{10})(n_{01})}{N^2};$$

the MBC (Reckase, 1997),

$$\mathrm{MBC}_{jj'} = \frac{\sum_{i=1}^{N}(X_{ij} - E(X_{ij}))(X_{ij'} - E(X_{ij'}))}{N};$$

and the closely related Q_3 (Yen, 1993),

$$Q_{3_{jj'}} = r_{e_{ij}e_{ij'}},$$

where r refers to the correlation and $e_{ij} = X_{ij} - E(X_{ij})$ where $E(X_{ij})$ is given by the item response function (Equation 1). The residual item covariance (Fu et al., 2005; McDonald & Mok, 1995) is given by

$$\mathrm{RESIDCOV}_{jj'} = \frac{[(n_{11})(n_{00}) - (n_{10})(n_{01})]}{N^2} - \frac{[E(n_{11})E(n_{00}) - E(n_{10})E(n_{01})]}{E(N^2)}.$$

Three other measures that are nonlinearly based measures of association are the natural log of the odds ratio (Agresti, 2002),

$$\mathrm{LN}(\mathrm{OR}_{jj'}) = \ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right] = \ln(n_{11}) + \ln(n_{00}) - \ln(n_{10}) - \ln(n_{01});$$

a standardized log odds ratio residual (Chen & Thissen, 1997),

$$\mathrm{STDLN}(\mathrm{OR}_{jj'}) - \mathrm{RESID} = \frac{\ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right] - \ln\left[\frac{E(n_{11})E(n_{00})}{E(n_{10})E(n_{01})}\right]}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}};$$

and a Mantel-Haenszel statistic (MH; Agresti, 2002; Sinharay et al., 2006),

$$\mathrm{MH}_{jj'} = \frac{\sum_r n_{11r} n_{00r}/n_r}{\sum_r n_{10r} n_{01r}/n_r},$$

where $n_{11}r$ refers to the number of examinees with rest score r with a response pattern of $j = 1$ and $j' = 1$, where the rest score r is defined as the total test score excluding items j and j' and $n_{10}r$, $n_{01}r$, and $n_{00}r$ are defined analogously.

In a PPMC environment, Sinharay and Johnson (2003) found the odds ratio to be a useful discrepancy measure for performing model criticism in a number of situations that result in local dependencies among items. Sinharay et al. (2006) expanded on that work and found the MH statistic to be an even more effective measure. McDonald and Mok (1995, in a frequentist framework) and Fu et al. (2005, in a PPMC framework) recommended the use of residual item covariances. Chen and Thissen (1997) found Q_3 to be an effective measure for detecting local dependence but found that it did not exhibit the assumed $N(0,\ 1)$ distribution under null conditions. This limitation is overcome in a PPMC framework, as the reference distribution is not assumed, but rather constructed empirically (Sinharay et al., 2006).

Investigating the use of PPMC to detect compensatory multidimensionality, Levy et al. (2009) found that (a) the bivariate X^2 and G^2 discrepancy measures were insensitive to the direction of the local dependence induced by the multidimensionality (i.e., they could not distinguish between item pairs with negative local dependence and item pairs with positive local dependence); (b) the covariance, residual covariance, log odds ratio, and the standardized log odds ratio performed similarly to each other; and (c) MBC and Q_3 performed almost identically and outperformed the other discrepancy measures except for MH which held a slight edge under null and non-null conditions. Under null conditions of unidimensionality, the distributions of PPP values for MBC, Q_3, and MH were nearly uniform, indicating near-optimality. The implication is that the use of PPMC with these measures in a hypothesis testing framework yields empirical Type I error rates quite close to nominal values. Of key interest in the present study is whether the performance of these measures in the current context of conjunctive MIRT mirrors their performance in the compensatory case.

In addition, three univariate discrepancy measures were investigated at the item level: the proportion correct and univariate X^2 and G^2 measures. These measures were found to be completely ineffective in detecting the presence of multidimensionality and as such will not be discussed in further. These results complement the findings in the analysis of compensatory multidimensional data and further support the hypothesis that univariate measures are poorly suited to address issues of dimensionality (Levy et al., 2009).

Results

For each data set within a condition, each discrepancy is evaluated once for each unique pairing of the 32 items, resulting in 496 PPP values. Four types of item pairs were defined: pairs in which (a) both items reflect θ_1 only, (b) one item reflects θ_1 only and one item reflects θ_1 and one of the auxiliary dimensions (either θ_2 or θ_3), (c) both items reflect θ_1 and the same auxiliary dimension, and (d) one item reflects θ_1 and θ_2 and one reflects θ_1 and θ_3 (i.e., the items in the pair reflect the primary and different auxiliary dimensions).[1]

Table 2 presents results for multivariate analyses of variance (MANOVA) for the discrepancy measures. Separate analyses were conducted for the different

TABLE 2

MANOVA Results for Bivariate Discrepancy Measures

Mult. Items	Source	Both Items Reflect θ_1					One Item Reflects θ_1 One Item Reflects θ_1 and Either θ_2 or θ_3					Both Items Reflect θ_1 and θ_2 or Both Items Reflect θ_1 and θ_3					One Item Reflects θ_1 and θ_2 One Item Reflects θ_1 and θ_3				
		Wilks's Λ	F	df 1	df 2	η^2	Wilks's Λ	F	df 1	df 2	η^2	Wilks's Λ	F	df 1	df 2	η^2	Wilks's Λ	F	df 1	df 2	η^2
4	Strength (S)	1.000	3.56	24	2628616	0.000	0.830	2158.41	24	778204	0.060	0.673	83.90	24	13763	0.124	0.585	234.27	24	27681	0.164
	Correlation (ρ)	1.000	12.40	24	2628616	0.000	0.836	2067.43	24	778204	0.058	0.681	81.33	24	13763	0.120	0.626	202.38	24	27681	0.145
	Sample size (N)	0.983	959.20	16	1812648	0.008	0.983	291.46	16	536636	0.009	0.774	81.16	16	9490	0.120	0.928	45.59	16	19088	0.037
	S × ρ	1.000	2.32	72	5512961	0.000	0.890	438.27	72	1632122	0.014	0.959	2.79	72	28870	0.005	0.815	27.61	72	58061	0.025
	S × N	1.000	2.06	48	4459498	0.000	0.987	70.90	48	1320241	0.002	0.882	12.62	48	23351	0.021	0.943	11.77	48	46965	0.010
	ρ × N	1.000	2.28	48	4459498	0.000	0.990	55.54	48	1320241	0.002	0.796	23.12	48	23351	0.037	0.926	15.42	48	46965	0.013
	S × ρ × N	1.000	2.11	144	6668153	0.000	0.997	6.22	144	1974122	0.000	0.919	2.81	144	34924	0.011	0.950	3.42	144	70232	0.006
16	Strength (S)	0.997	33.91	24	834041	0.001	0.917	2240.64	24	1776826	0.028	0.861	856.67	24	389451	0.049	0.800	1481.59	24	444919	0.072
	Correlation (ρ)	0.977	274.59	24	834041	0.008	0.877	3413.67	24	1776826	0.043	0.791	1366.54	24	389451	0.075	0.762	1824.37	24	444919	0.087
	Sample size (N)	0.962	696.11	16	575140	0.019	0.975	965.11	16	1225268	0.012	0.875	1155.76	16	268558	0.064	0.953	465.78	16	306808	0.024
	S × ρ	0.998	7.38	72	1749228	0.000	0.948	454.78	72	3726515	0.007	0.985	28.57	72	816795	0.002	0.918	183.25	72	933128	0.011
	S × N	0.999	3.57	48	1414969	0.000	0.992	107.19	48	3014420	0.001	0.965	101.30	48	660713	0.006	0.974	83.80	48	754816	0.004
	ρ × N	0.995	30.49	48	1414969	0.001	0.994	77.93	48	3014420	0.001	0.917	245.48	48	660713	0.014	0.962	123.40	48	754816	0.006
	S × ρ × N	0.999	2.96	144	2115766	0.000	0.998	8.35	144	4507374	0.000	0.982	16.90	144	987951	0.002	0.985	16.14	144	1128660	0.002
28	Strength (S)	0.952	29.06	24	40988	0.016	0.977	258.42	24	773738	0.008	0.892	2112.82	24	1266311	0.037	0.848	3320.68	24	1363190	0.053
	Correlation (ρ)	0.865	87.91	24	40988	0.047	0.963	427.62	24	773738	0.013	0.868	2635.84	24	1266311	0.046	0.761	5618.02	24	1363190	0.087
	Sample size (N)	0.847	152.96	16	28264	0.080	0.958	720.77	16	533556	0.021	0.925	2154.58	16	873226	0.038	0.921	2457.01	16	940032	0.040
	S × ρ	0.974	5.28	72	85969	0.003	0.988	45.56	72	1622755	0.002	0.991	51.94	72	2655821	0.001	0.941	396.68	72	2859003	0.008
	S × N	0.987	3.92	48	69539	0.002	0.994	33.99	48	1312664	0.001	0.984	143.27	48	2148323	0.003	0.981	186.80	48	2312680	0.003
	ρ × N	0.966	10.14	48	69539	0.006	0.997	16.65	48	1312664	0.000	0.948	485.58	48	2148323	0.009	0.952	488.43	48	2312680	0.008
	S × ρ × N	0.984	1.58	144	103988	0.002	0.999	2.70	144	1962792	0.000	0.990	29.68	144	3212327	0.001	0.988	37.91	144	3458084	0.001

Note: Analyses were conducted separately based on the type of item pair and number of multidimensional items. For all test statistics, $p < .001$.

types of item pairs and different numbers of multidimensional items. All tests were significant at the .001 level; inspection of the values of the η^2 reveals the relative importance of the manipulated factors. Broadly speaking, the results indicate that the manipulated factors have larger effects on item pairs in which both items are multidimensional and the relative lack of importance of the interaction terms.

To facilitate finer-grained inferences about the effects of the manipulated factors and pursue differences in the performances of the discrepancy measures, graphical representations are presented. The results for the large sample size are presented first in terms of median PPP values followed by proportions of extreme PPP values. Analogous results for the remaining sample sizes will not be presented, as the patterns at the small and moderate sample sizes mimicked those at the large sample size. The effects of sample size will be incorporated into the presentation of the proportion of extreme PPP values. Figure 1 plots the median PPP values for the X^2 discrepancy measure for pairs of items for the conditions with a sample size of 2,500. There are 16 panels in the plot, corresponding to the combinations of the four levels of strength of dependence (i.e., the value of b_{j2} and b_{j3}) with the four levels of the correlations among the latent variables. Within each panel, the horizontal axis is the number of items influenced by the second or third dimension (4, 16, or 28). The height of each point gives the value of the median PPP value, separate for each of the four types of item pairs. The medians were computed after pooling over each instantiation of each type of item pair and over the 50 replications within the condition.[2]

Figure 2 plots the median PPP values for MBC. The structure of the plots is identical to that of Figure 1. Plots for the remaining bivariate discrepancy measures will not be presented on space considerations. The patterns for the G^2 discrepancy measure for item pairs mimicked those for the X^2 discrepancy measure for item pairs (Figure 1). The patterns for the remaining discrepancy measures (i.e., the log odds ratio, covariance, Q_3, residual covariance, standardized log odds ratio, and MH) mimicked those for MBC (Figure 2).

Though useful for tracking the general behavior of the PPP values across the conditions, the median is not optimal for summarizing the extreme values that lie in the tails. To that end we computed the rates at which the discrepancy measures yielded extreme PPP values. Operationally, we consider a PPP value to be extreme when it is less than .05 or greater than .95. In a hypothesis testing framework, the proportions of extreme PPP values are power rates based on a two-tailed test with $\alpha = .10$. The results for the proportion of extreme PPP values for X^2, the log odds ratio, MBC, and MH are presented and discussed. Owing to the similarity of the results for some of the measures, presenting these measures is sufficient to summarize all the results. The results for X^2 are representative of G^2; the results for MBC are representative of Q_3. The results for the log odds ratio are representative of the covariance, residual covariance, and the standardized log odds ratio residual.

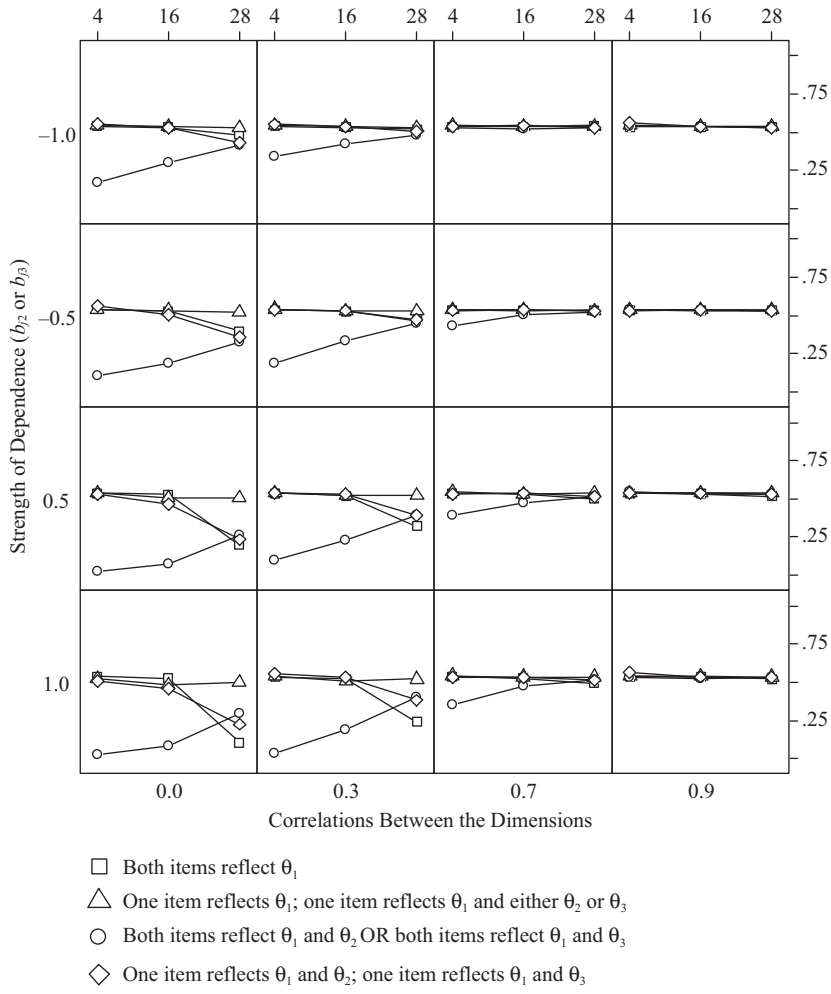

FIGURE 1. *Median PPP values for* X^2 *for item pairs when* $N = 2{,}500$. *Within panels the horizontal axis is the number of items influenced by the second or third dimension. Results are also representative of those for the* G^2 *discrepancy measure for item pairs.*

The panels in Figure 3 plot the proportion of extreme PPP values for item pairs that reflect the same multiple dimensions. Figures 4 and 5 plot the proportions for item pairs that reflect different multiple dimensions and for item pairs in which both items reflect the primary dimension only (respectively). The results for the remaining type of item pairs, in which one item reflects the primary dimension only and the other item reflects multiple dimensions will not be presented, as the medians for this type of item pair (i.e., the triangles in Figures 1 and 2) did not

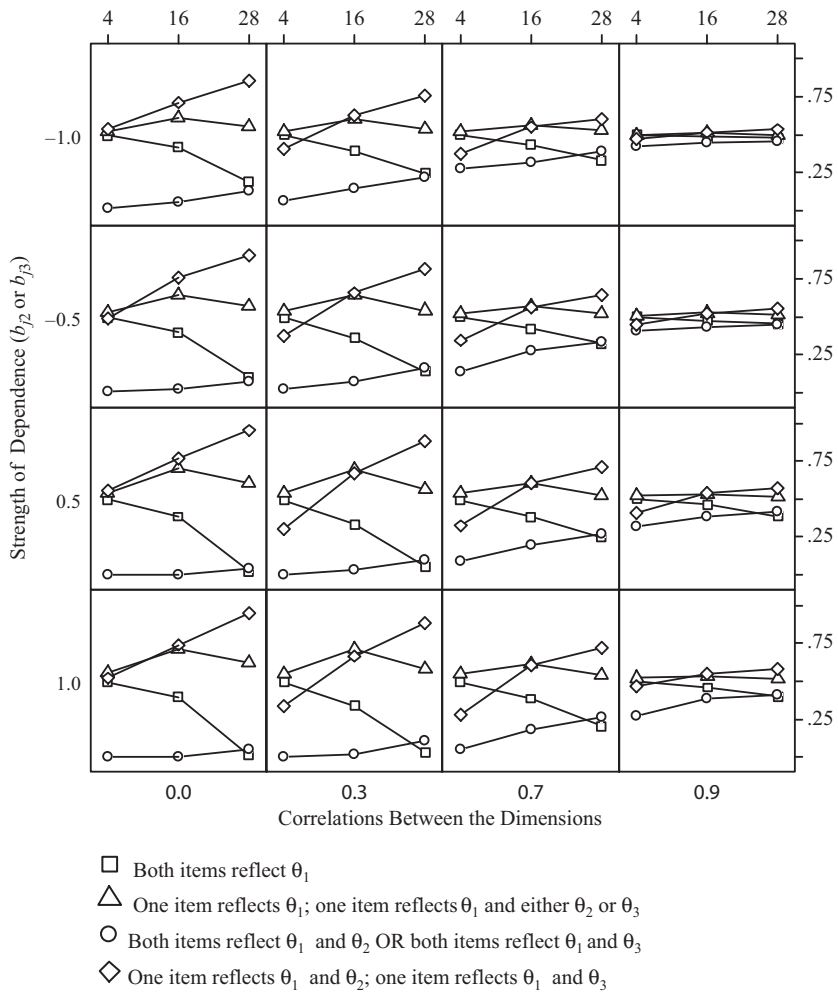

FIGURE 2. *Median PPP values for MBC when* N = *2,500. Within panels the horizontal axis is the number of items influenced by the second or third dimension. Results are also representative of those for the log odds ratio, covariance,* Q^3*, residual covariance, standardized log odds ratio, and MH.*

meaningfully deviate from .5 for any of the discrepancy measures. The proportions of extreme PPP values for these item pairs are quite low and do not show a systematic pattern.

Figure 6 plots the proportion of extreme PPP values for MH for item pairs that reflect the same multiple dimensions across the sample sizes. Within each panel the three sets of points (squares, triangles, and circles) correspond to conditions

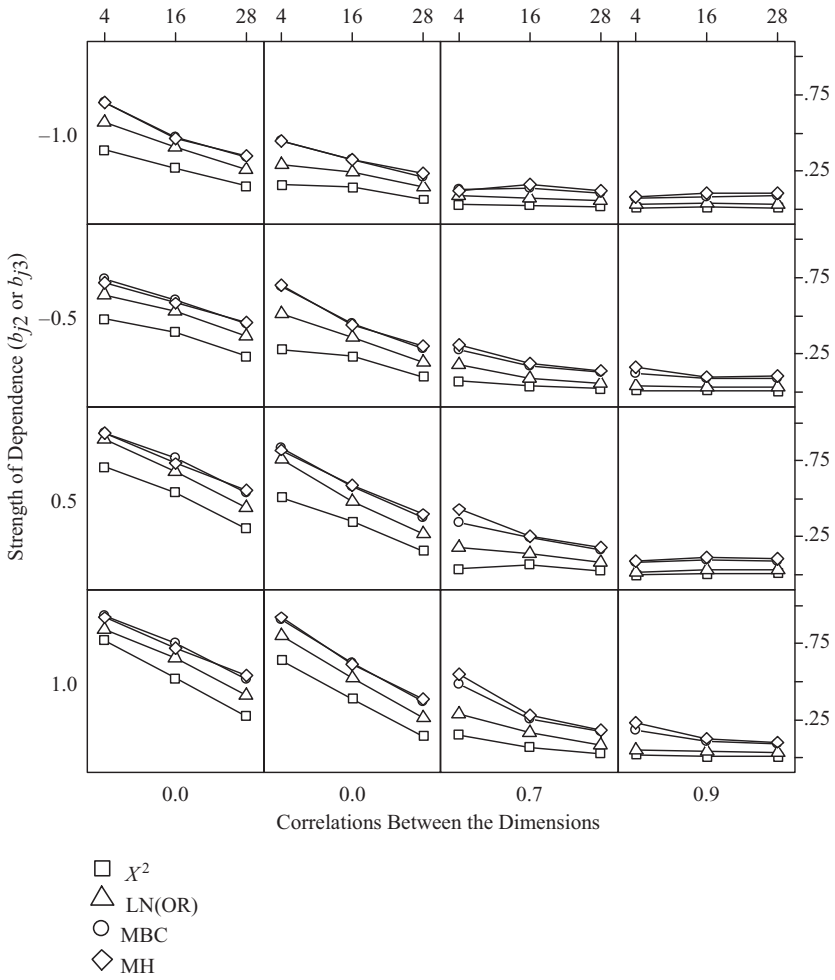

FIGURE 3. *Proportion of extreme PPP values (i.e., PPP value < .05 or >.95) for select discrepancy measures for item pairs that reflect the same multiple dimensions when* N = *2,500. Within panels, the horizontal axis is the number of items influenced by the second or third dimension.*

with 4, 16, and 28 multidimensional items. Progressing left to right within a panel, the points plot the rates of exhibiting extreme PPP values as sample size increases from 250 to 750 and to 2,500. The remaining bivariate discrepancy measures and types of item pairs exhibited the same overall pattern (i.e., increases in sample sizes yielded higher proportions of extreme PPP values) and are not displayed because of space considerations.

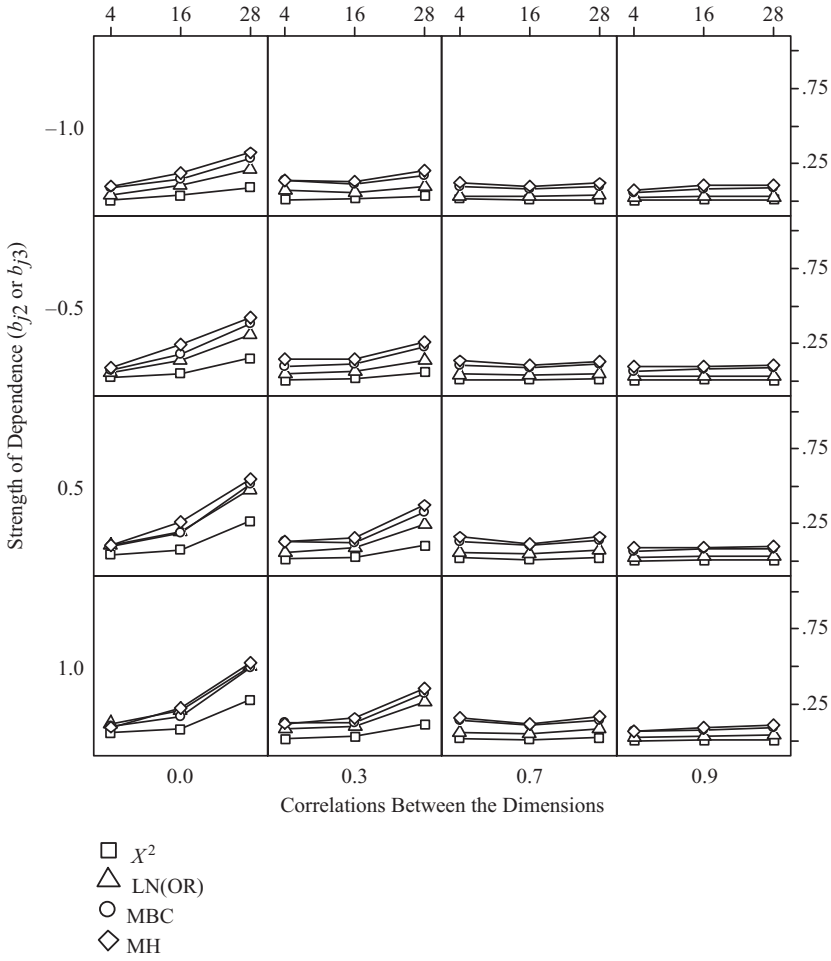

FIGURE 4. *Proportion of extreme PPP values (i.e., PPP value < .05 or >.95) for select discrepancy measures for item pairs that reflect different multiple dimensions when* N = *2,500. Within panels, the horizontal axis is the number of items influenced by the second or third dimension.*

Discussion of the Study

The results reveal that the performance of PPMC for detecting the multidimensionality improves (i.e., the PPP values become more extreme) as (a) the strength of dependence on auxiliary dimensions increases (Figures 1–6), (b) the correlations between the latent dimensions decrease (Figures 1–6), and (c) sample size increases (Figure 6). These patterns for the strength of dependence and correlations among the

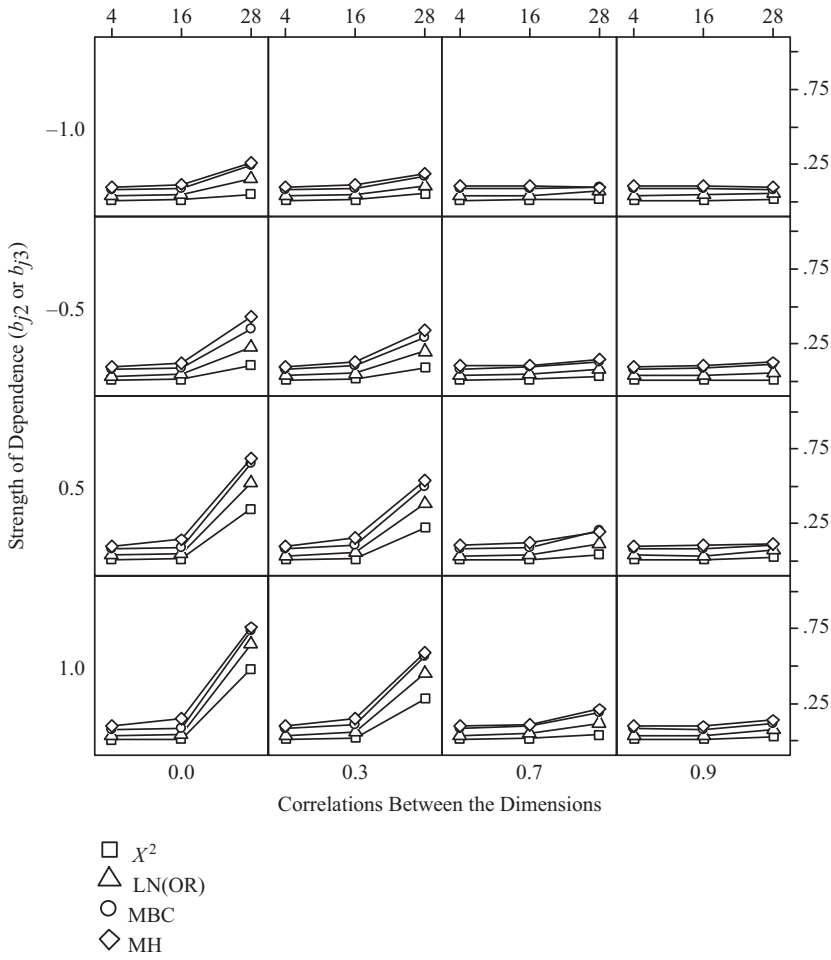

FIGURE 5. *Proportion of extreme PPP values (i.e., PPP value < .05 or >.95) for select discrepancy measures for item pairs that reflect the primary dimension only when* N = *2,500. Within panels, the horizontal axis is the number of items influenced by the second or third dimension.*

latent dimensions were observed for all types of item pairs except pairings of items in which one item reflects the primary dimension only and the other item reflects multiple dimensions. Note, however, that the principal effects of the manipulated factors were not present in all combinations of the conditions. For example, when the correlations between the dimensions were extremely strong (.9), the remaining factors became almost irrelevant. Even at the strongest levels of dependence and the largest sample size, it became nearly impossible to detect the multidimensionality.

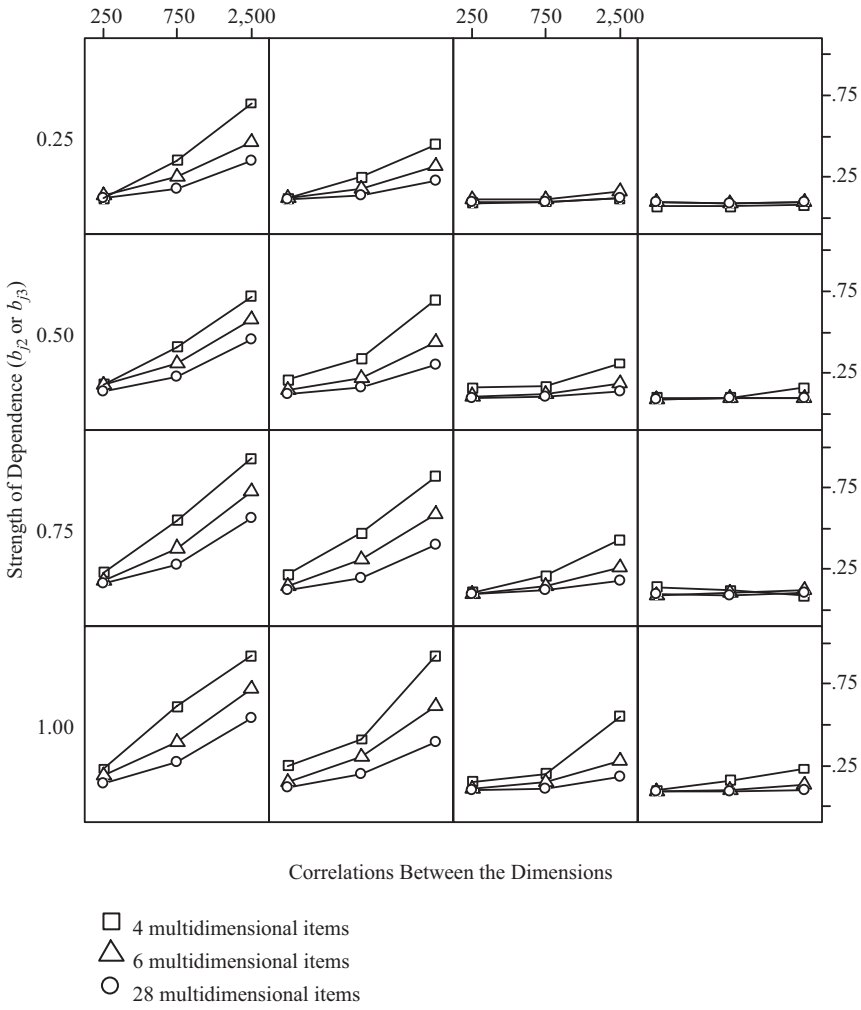

FIGURE 6. *Proportion of extreme PPP values (i.e., PPP value < .05 or >.95) for the MH for item pairs that reflect the same multiple dimensions. Within panels, the horizontal axis lists the sample size.*

For item pairs that reflect the same auxiliary dimension, the proportions of extreme PPP values decrease as the proportion of multidimensional items increases (Figure 3). In contrast, the proportions for (a) item pairs that reflect different multiple dimensions (Figure 4) and (b) item pairs that reflect the primary dimension only (Figure 5) increase as the proportion of multidimensional items increases. This pattern is also evident in the analysis of the median PPP

values. In Figures 1 and 2, it is observed that the medians for item pairs that reflect the same auxiliary dimension (depicted as circles) get closer to .5 as the proportion of multidimensional items increases, whereas the medians for item pairs that reflect different multiple dimensions or the primary dimension only (depicted as diamonds and squares, respectively) mover farther away from .5.

These results imply that when relatively few items depend on multiple dimensions, the multidimensionality will manifest itself most prominently in terms of item pairs that reflect the same multiple dimensions (Figure 3). At high proportions of multidimensional items, it becomes harder to detect the multidimensionality in terms of these pairings of items, but easier to detect the multidimensionality in terms of item pairs that reflect different multiple dimensions or item pairs that reflect the primary dimension only (Figures 4 and 5).

The findings of the influences of the (a) strength of dependence of the items on auxiliary dimensions, (b) correlations among the dimensions, (c) proportion of multidimensional items, and (d) sample size closely mirror those found by Levy et al. (2009) in the context of compensatory multidimensional data, which was couched in the framework of conditional covariance theory for compensatory multidimensional data. However, the conjunctive multidimensional IRT models considered here do not belong to the class of generalized compensatory models to which that theoretical framework directly applies (Zhang & Stout, 1999). The results of the current work suggest that these factors do indeed influence dimensionality assessment in the conjunctive multidimensional context.

Key differences were observed in terms of the discrepancy measures themselves. As hypothesized, most of the discrepancy measures yielded increasingly large PPP values for item pairs that reflect different dimensions as the proportion of multidimensional items increases (i.e., the diamonds in each panel of Figure 2 start close to .5 and increase as the proportion increases). However, the PPP values for the X^2 and G^2 measures for these item pairs decreased (see the diamonds in Figure 1), a result of these measures being nondirectional (Chen & Thissen, 1997) and therefore incapable of distinguishing between positive and negative local dependence (Levy et al., 2009). The covariance, log odds ratio, MBC, Q_3, residual covariance, standardized log odds ratio residual, and MH are sensitive to the directionality of misfit and may be more useful in substantive interpretations of statistical analyses of data-model fit for model criticism and model reformulation (Levy et al., 2009).

Differences between some of the discrepancy measures were observed in terms of the magnitudes of the PPP values and the proportion of extreme PPP values (Figures 3–5). The X^2 and G^2 measures resulted in the lowest rates of extreme PPP values across the conditions. The covariance and residual covariance performed quite similarly to each another. Likewise, the log odds ratio and the standardized log odds ratio performed quite similarly to each other. What is more, these two groups (i.e., the covariance, residual covariance, log odds ratio, and standardized log odds ratio residual) performed similarly.

The most effective measures were MBC, Q_3, and MH. MBC and Q_3 performed nearly identically across the conditions; neither consistently outperformed the other, and the differences, when present, were trivial. MH performed similarly to these measures in many cases (as evidenced by the frequent overlapping of the circles and diamonds in Figures 3–5). In some cases, MH slightly outperformed MBC and Q_3.

The overall ordering of discrepancy measures in terms of performance can be summarized as follows: X^2 and G^2 were comparable to one another and the worst, followed by the covariance, residual covariance, log odds ratio, and standardized log odds ratio residual, which were all similar, followed by MBC, Q_3, and MH, which were the best. This relative ordering of performance—and the superiority of MBC, Q_3, and MH in particular—was also present in an analysis of unidimensional data and compensatory multidimensional data (Levy et al., 2009). As such, MBC, Q_3, and MH are recommended as measures for conducting PPMC to investigate the possibility of multidimensionality in the context of unidimensional modeling.

These conclusions must be viewed in light of an understanding of the links between the chosen model, potential data structures, and discrepancy measures. It is imperative that the analyst, guided by an understanding of potential inadequacies of the model for the data under consideration, purposefully select discrepancy measures that target different aspects of data-model agreement in order to diagnose the model's strengths and weaknesses. Choices for the discrepancy measures should be guided by jointly considering the data, model, and potential discrepancy measures. Several examples from the study illustrate this point. Consideration of the 2-PLM, which has a unique location parameter for every item, implies that the proportion correct will not be a useful discrepancy measure for any situation of model criticism (Levy et al., 2009). An understanding of the implications of multidimensionality for producing covariation between items suggests that univariate discrepancy measures will not be effective, but that bivariate discrepancy measures may be effective. Familiarity with the bivariate X^2 and G^2 discrepancy measures implies that they will be insensitive to the direction of local dependence (Chen & Thissen, 1997; Levy et al., 2009), whereas familiarity with the remaining bivariate measures implies that they will be sensitive to the direction of local dependence. What is more, a conditional covariance theory perspective suggests the utility of discrepancy measures that condition on proficiency when examining inter-item associations (Stout et al., 1996; Zhang & Stout, 1999). In the current work, MBC, Q_3, and MH (where the latter employs the rest score as a proxy for proficiency) are such measures and were found to be the most successful in the current study on conjunctive multidimensionality as well as in other studies of unidimensionality and compensatory multidimensionality (Levy et al., 2009).

Concluding Remarks

The current work extends the investigations of PPMC applied to IRT modeling to the case of critiquing unidimensional models fit to conjunctive

MIRT data. Key findings regarding the manipulated factors are that the (a) relative strength of dependence of the items on the latent dimensions, (b) correlations among dimensions, (c) proportion of multidimensional items, and (d) sample size all influence the ability to detect multidimensionality. The findings in the context of conjunctive MIRT data structures point to the generality of these influences, which were initially motivated by geometric representations for linear relations of compensatory models (Levy et al., 2009). The current study presents empirical evidence in the context of conjunctive multidimensional structures that support the relevance of these influences as being more general than their compensatory manifestations.

Results of the analysis of the different discrepancy measures include the (a) ineffectiveness of univariate measures, (b) insensitivity of the bivariate X^2 and G^2 discrepancy measures to the direction of misfit, and (c) superiority of MBC, Q_3, and MH. These findings complement earlier work on situations of compensatory multidimensionality as well as null conditions of unidimensionality (Levy et al., 2009).

A number of different aspects of this work are deserving of further attention. The hypothesis that the (a) relative strength of dependence of the items on the latent dimensions, (b) correlations among dimensions, (c) proportion of multidimensional items are general factors that influence dimensionality assessment can be explored in other contexts. In terms of the behavior of the PPMC, further research on MBC, Q_3, and MH is needed to fully explore their potential, particularly in consideration of additional discrepancy measures shown to be useful in related contexts (e.g., Hoijtink, 2001; Sinharay et al., 2006).

Still other extensions include conducting PPMC to criticize more complex models. Zhang and Stout (1999) characterized approaches to dimensionality assessment in terms of those that (a) are more exploratory in nature, and attempt full dimensionality assessment by estimating the number of latent dimensions and determining which items reflect which dimension, or (b) are more confirmatory in that they assess unidimensionality. PPMC fits into neither of these categories. PPMC is confirmatory in nature but not restricted to the assessment of unidimensionality. The flexibility of PPMC may be leveraged to provide more general confirmatory dimensionality assessment and model criticism tools to be applied when the model is unidimensional or multidimensional. The current work evidences the utility of PPMC for criticizing unidimensional models and is suggestive of the potential of PPMC for more complex situations.

Notes

1. In some cases the G^2, log odds ratio, standardized log odds ratio residual, and MH could not be computed due to zero frequencies for counts of bivariate

response patterns (Chen & Thissen, 1997). This was quite rare, occurring in no more than 0.43% of the computations for any one condition. These cases were ignored from the analyses.

2. The utility of pooling all the instantiations of item pairs of a given type (within each condition) was investigated by considering the instantiations separately. These analyses indicated that though the PPP values for the instantiations of each type of item pair varied in terms of magnitude, they were quite consistent in their direction relative to .5. As such, a discussion at this disaggregated level would not differ substantively, in terms of the overall patterns and trends, from that which is presented here.

Acknowledgment

This research was supported in part by an ETS Harold Gulliksen Psychometric Research Fellowship. The author would like to express his sincere thanks to Bob Mislevy and Sandip Sinharay for guidance on this work and comments on earlier drafts of this article. The author would also like to thank Carl Lejuez, Karen Samuelsen, Marc Kroopnick, and Daisy Rutstein for their generosity with regard to computational resources. The author would also like to thank the editor and reviewers for comments that led to improvements over previous versions of the article.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York, NY: Wiley.

Berkhof, J., van Mechelen, I., & Gelman, A. (2004). *Enhancing the performance of a posterior predictive check (IAP Statistics Network Technical Report 0350).* Louvain-la-Neuve: Belgium, Interuniversity Attraction Pole, Katholieke Universiteit Leuven.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika*, *23*, 67–95.

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.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, *49*, 175–186.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York, NY: Springer-Verlag.

Fu, J., Bolt, D. M., & Li, Y. (2005, April). *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, Montréal, Canada.

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

Hoijtink, H. (2001). Conditional independence and differential item functioning in the two-parameter logistic model. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays in item response theory* (pp. 109–130). New York, NY: Springer-Verlag.

Jannsen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, *25*, 285–306.

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

McDonald, R. P., & Mok, M. M.-C. (1995). Goodness of fit in item response models. *Multivariate Behavioral Research*, *30*, 23–40.

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

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.

Reckase, M. D. (1997). A linear logistic multidimensional model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York, NY: Springer-Verlag.

Robins, J. M., van der Vaart, A., & Ventura, V. (2000). The asymptotic distribution of P values in composite null models. *Journal of the American Statistical Association*, *95*, 1143–1172.

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

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. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models (ETS research report RR-03–28).* Princeton, NJ: Educational Testing Service.

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

Stern, H. S. (2000). Comment. *Journal of the American Statistical Association*, *95*, 1157–1159.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331–354.

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

Zenisky, A. L., Hambleton, R. K., & Sireci, S. G. (2003). *Effects of local item dependence on the validity of IRT item, test, and ability statistics* [MCAT Monograph no. 5].

Zhang, J., & Stout, W. (1999). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika*, *64*, 129–152.

Authors

ROY LEVY is an assistant professor at Arizona State University, PO Box 873701, Tempe, AZ, 85287-3701; e-mail: Roy.Levy@asu.edu. His research interests include methodological investigations and applications in item response theory, structural equation modeling, and Bayesian networks.

Manuscript received October 11, 2007
Revision received June 26, 2008
Accepted June 27, 2008