ABSTRACT

Title of dissertation: POSTERIOR PREDICTIVE MODEL CHECKING FOR MULTIDIMENSIONALITY IN ITEM RESPONSE THEORY AND BAYESIAN NETWORKS

Roy Levy, Doctor of Philosophy, 2006

Dissertation directed by: Professor Robert J. Mislevy
Department of Measurement, Statistics & Evaluation

If data exhibit a dimensional structure more complex than what is assumed, key conditional independence assumptions of the hypothesized model do not hold. The current work pursues posterior predictive model checking, a flexible family of Bayesian model checking procedures, as a tool for criticizing models in light of inadequately modeled dimensional structure. Factors hypothesized to influence dimensionality and dimensionality assessment are couched in conditional covariance theory and conveyed via geometric representations of multidimensionality. These factors and their hypothesized effects motivate a simulation study that investigates posterior predictive model checking in the context of item response theory for dichotomous observables. A unidimensional model is fit to data that follow compensatory or conjunctive multidimensional item response models to assess the utility of conducting posterior predictive model checking. Discrepancy measures are formulated at the level of

individual items and pairs of items. A second study draws from the results of the first study and investigates the model checking techniques in the context of multidimensional Bayesian networks with inhibitory effects. Key findings include support for the hypothesized effects of the manipulated factors with regard to their influence on dimensionality assessment and the superiority of certain discrepancy measures for conducting posterior predictive model checking on dimensionality assessment. The application of these techniques to models both familiar to assessment and those that have not yet become standard practice speaks to the generality of the procedures and its potentially broad applicability.

POSTERIOR PREDICTIVE MODEL CHECKING FOR MULTIDIMENSIONALITY IN ITEM RESPONSE THEORY AND BAYESIAN NETWORKS

by

Roy Levy

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2006

Advisory Committee:
Professor Robert J. Mislevy, Chair
Professor Robert W. Lissitz
Professor Gregory R. Hancock
Associate Professor Robert G. Croninger
Assistant Professor Amy B. Hendrickson

DEDICATION

To Paige, whose love has made this all worthwhile, and whose patience and support have made this all possible.

ACKNOWLEDGEMENTS

One of the joys of my work is that very little of it is done in complete isolation and this is no exception. A number of people and parties have made this work possible and rewarding. This research was supported in part by Educational Testing Service via a Harold Gulliksen Psychometric Research Fellowship. I thank ETS and in particular Dr. Sandip Sinharay, whose work was a source of inspiration and whose guidance has been instrumental in the genesis, maturation, and completion of this project.

I wish to thank Dr. Carl Lejuez, Dr. Karen Samuelsen, Marc Kroopnick, and Daisy Wise for their generosity with regard to computational resources. I also thank the avuncular Dr. Gregory R. Hancock, whose work, style, and advice has had a greater influence over this work than he knows. My deepest gratitude goes to Dr. Robert J. Mislevy, whose vision, patience, and encouragement have allowed me to explore my interests and whose wisdom has been and continues to be a guiding light when I find myself in need of direction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: PURPOSE AND RATIONALE

Statistical models used in educational assessment and psychological measurement are often constructed, viewed, and judged in terms of what they state exists (e.g., there is a common construct, this factor causes another factor). In terms of evaluating a model, it may be preferable to consider a model as stating what does *not* exist, that is, in terms of what restrictions the model imposes (Mulaik, 2001). Models have implications for how observed data ought to behave. Data-model fit assessment and model comparison procedures come to characterizing the discrepancy between observed and implied behavior of data.

CONDITIONAL INDEPENDENCE IN PSYCHOMETRIC MODELS

Most psychometric models in use assume a construct or set of constructs underlying performance on observable variables. A hallmark of modern statistical psychometric models is the use of latent variables in modeling such situations. Examples include item response theory (IRT; Hambleton & Swaminathan, 1985), factor analysis and other members of the structural equation modeling (SEM) family (Bollen, 1989), and latent class analysis (LCA; Dayton, 1998), which differ in their assumptions about the properties of the variables analyzed and often in their purposes (e.g., evaluation of relationships between constructs, scaling of items, measurement of individuals, etc.).

An important class of constraints for these models involves conditional independence assumptions. For $i = 1, \ldots, N$, let $\boldsymbol{\theta}_i$ denote a possibly vectored-valued latent variable for subject *i* and let X_{ij}, $j = 1, \ldots, J$, denote an observable variable whose value corresponds to the scored response *j* from subject *i* to some task or test item

(allowing for multivariate scores). Variables characterizing the psychometric properties of observable j (e.g., item parameters, factor loadings) are contained in a vector, denoted $\boldsymbol{\omega}_j$. A psychometric model typically structures the joint distribution of the observables by positing that the value of any observed variable is conditionally independent of all other variables, given the values of the subject's latent variables contained in $\boldsymbol{\theta}_i$ and the variables defining the distributional properties of the observable, $\boldsymbol{\omega}_j$:

$$P\left(X_{ij} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_J, X_{11}, \ldots, X_{NJ}\right) = P\left(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j\right). \quad (1)$$

The conditional independence of any other subject's variables (latent or observed) is termed *respondent independence*. Within subjects, the conditional independence of an observable of any other observable is termed *local independence*. An important distinction is the difference between strong and weak versions of local independence (e.g., McDonald, 1997; Stout et al., 1996). Strong local independence states that a subject's value on one observable is independent of all other observables, conditional on the latent variable. Let $\mathbf{X}_i$ denote the vector of observables from subject *i*. The joint distribution of observables for any subject is then the product of the univariate distributions,

$$P\left(\mathbf{X}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}\right) = \prod_{j=1}^{J} P\left(X_{ij} \mid \boldsymbol{\theta}_i, \boldsymbol{\omega}_j\right).$$

Weak local independence states that each subject's observables are linearly independent. As a less stringent requirement, weak local independence is necessary but not sufficient for strong local independence (Stout et al., 1996).

In practice, weak local independence is most often investigated in terms of pairs of observables as it is reasonable for analysts to assume that if variables are pair-wise

independent, higher order dependencies, though possible, are highly implausible (McDonald, 1997; McDonald & Mok, 1995). This is often operationalized in terms of the conditional covariance between observables; if weak local independence holds,

$$\text{Cov}\left(X_j, X_{j'} \mid \boldsymbol{\theta}_i\right) = 0, \quad j' \neq j .$$

In terms of modeling and model checking, the conditional independence assumptions have implications for how data ought to behave. Data-model fit assessment comes to judgments about the discrepancy between the observed and model-implied behavior of data.

IRT, factor analysis, SEM, LCA, and other families of models differ in their assumptions regarding the latent and observed variables (Bartholomew, 1987). The preceding discussion of conditional independence and many of the multidimensional models for psychometric phenomena cut across models of different types, including classical test theory models that do not formally posit latent variables (Yen, 1993). Hence, investigations of conditional independence assumptions and these phenomena have potentially broad applications. The current work focuses attention on (a) cumulative, probabilistic IRT models that assume a single latent continuous dimension and a set of observable dichotomous variables and (b) cumulative, probabilistic models with IRT and LCA components and complex relationships. The former are among the most prevalent in current assessment practice; background for these models as it relates to the current work is reviewed in the next sections. The latter represent recent developments in modeling innovative assessments; background for these models is reviewed as part of a larger discussion of Bayesian networks.

UNIDIMENSIONAL LATENT VARIABLE MODELS

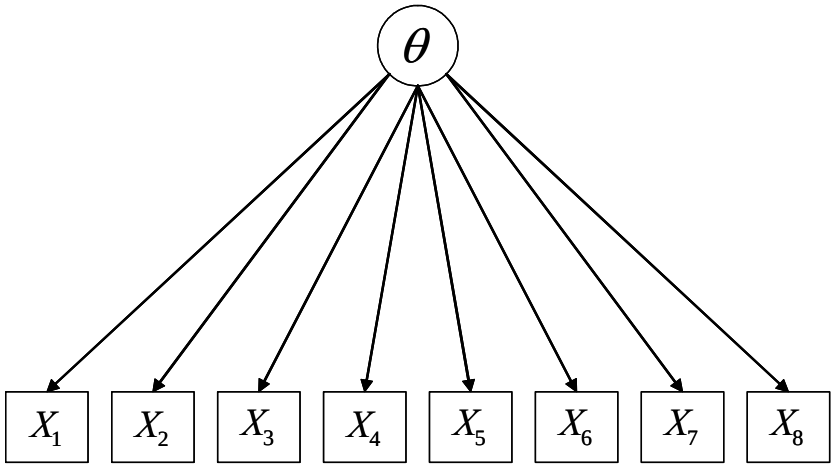
This section reviews the structure of unidimensional latent variable models, which are common in operational assessment. Though the focus of this work is on IRT and related Bayesian networks, the following characterization of unidimensional models is broader than one couched in IRT. This breadth is intentional, as issues of dimensionality cut across the particulars of different psychometric models. This characterization of unidimensionality and multidimensionality supports the generalizability of the results of this work to settings such as factor analysis and LCA that differ in their surface features but share a common basis in terms of dimensionality and the relationships between latent and observed variables.

A unidimensional latent variable model specifies that a single latent variable underlies the values of a multivariate distribution in the sense that the single latent variable is modeled as the lone source of association for the variables. Figure 1 depicts a unidimensional model in which a single latent dimension, θ, underlies eight observable variables, $X_1, \ldots, X_8$. Following common path diagrammatic notation (e.g., Bollen, 1989), we employ circles to represent latent variables, squares to represent observable variables, and the directed arrows to indicate a dependence of the variable at the head of the arrow on the variable at the tail.

In an assessment context, such models are appropriate when a single dimension or construct is of interest. In measurement terms, the model specifies that there is a latent variable of inferential interest and that there are multiple observed variables that are viewed as imperfect measures that depend on or relate to the latent variable. For example, suppose the eight observables in Figure 1 correspond to the scored responses

from an eight item test of mathematics. The latent variable θ might then be interpreted as mathematical or quantitative proficiency.

Figure 1: A unidimensional model.

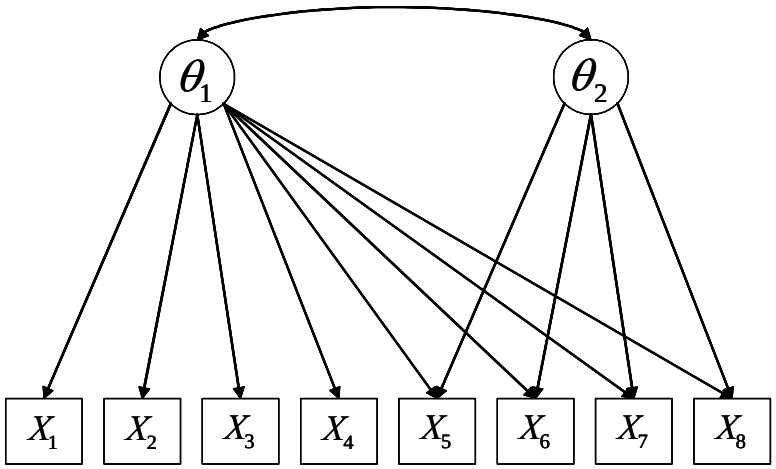


MULTIDIMENSIONAL LATENT VARIABLE MODELS

The focus of this work is on multidimensional models in which all observables reflect the primary dimension of inferential interest, but certain observables additionally reflect auxiliary dimensions. Figure 2 depicts this situation, in which all items reflect θ_1 and the last four items also reflect θ_2. Returning to the example, suppose the last four items on the math test are word problems, requiring the examinee to extract salient pieces of information from a textual passage. In this case, the latent dimensions θ_1 and θ_2 might then be interpreted as mathematics and reading proficiencies, respectively. The bi-directional arrow connecting θ_1 and θ_2 represents that the latent variables are permitted to be correlated, as might be expected. The multidimensional structure of the last four items indicates that performances on these items reflect *both* mathematics *and*

reading proficiencies. This type of structure has been variously termed within-item multidimensionality (Adams, Wilson, & Wang, 1997), heterogeneous test structure (Lucke, 2005), or factorially complex structure (McDonald, 2000).

Figure 2: A multidimensional model.



VIOLATIONS OF LOCAL INDEPENDENCE ASSUMPTIONS

A unidimensional model implies that the observed variables are conditionally independent (i.e., locally independent) given the latent variable. In Figure 1, $X_1,\ldots,X_8$ are related to each other because of their common dependence on θ. Conditioning on θ, $X_1,\ldots,X_8$ become independent, as there are no other sources of association for these variables.

If multidimensionality is present, multiple dimensions may be needed to render items conditionally independent. In Figure 2, conditioning on θ_1 renders $X_1,\ldots,X_4$ independent of one another and of the remaining observables. However, conditioning on

θ_1 is *not* sufficient to render $X_5, \ldots, X_8$ independent from one another, as there is an additional source of association, θ_2, that induces dependencies amongst $X_5, \ldots, X_8$. In order to render all the items conditionally independent, we must condition on both θ_1 and θ_2.

This underscores the point that local (in)dependence is best thought of as *with respect to* another variable or set of variables (Stout, 1987). For example, X_5 and X_6 in Figure 2 are locally *dependent* with respect to θ_1 but are locally *independent* with respect to θ_1 and θ_2. This frequently goes unarticulated. Owing to the primacy of unidimensional models, it is often the case that items are referred to as "locally independent" when a single dimension is sufficient to account for their association and "locally dependent" when they exhibit dependencies above and beyond that which can be accounted for by a single dimension. The preceding analysis of items that exhibit multidimensionality suggests that deeming items to be locally (in)dependent may be ambiguous as items may be locally dependent with respect to (i.e., when conditioning on) certain variables but locally independent with respect to (conditioning on) other variables. In the balance of this work, statements of local (in)dependence will be meant with respect to the assumed model.

If examinees differ on the multiple dimensions that influence performance on a set of items, a unidimensional model is inappropriate and a multidimensional model should be employed (Ackerman, 1994). Common assessment phenomena including differential item functioning, item drift, testlet effects, learning during testing, rater effects, and method effects, if unaccounted for, constitute violations of local independence and can be framed in terms of multidimensionality (e.g., Bolt & Stout,

1996; Bradlow, Wainer, & Wang, 1999, Mellenbergh, 1994; Shealy & Stout, 1993; Stout et al., 1996; Yen, 1993; Zwinderman, 1997).

Though many situations that result in a lack of local independence are often associated with unintended and undesirable phenomena, there are settings in which including dependent items are desirable. A simple example is the use of a common stimulus (e.g., reading passage) for a set of items. Many real world tasks involve simultaneously or sequentially addressing related problems; the inclusion of dependent tasks may therefore contribute to construct validity (Yen, 1993; Zenisky, Hambleton, & Sireci, 2003). Moreover, complex assessments in which tasks are designed to reflect multiple, possibly related traits exhibit multidimensionality by construction (Levy & Mislevy, 2004). The reader is referred to Yen (1993) for a comprehensive list of causes of local dependence.

If the data exhibit multidimensionality, the local independence assumptions implied by the use of a unidimensional model do not hold. Incorrectly assuming that observables are conditionally independent may lead to incorrect estimates of the values of variables as well as the precision of the estimates, which may have deleterious effects on (a) estimates of information and precision, (b) test assembly for target test information functions, (c) task selection and stopping rules in adaptive testing, (d) equating and linking, (e) inferences and decisions based on estimates, and (f) reporting estimates and the precisions of estimates (Birnbaum, 1968; Bradlow et al., 1999; Chen & Thissen, 1997; Mislevy & Patz, 1995; van der Linden, 1996; Yen, 1993).

As such, investigations of local independence assumptions are crucial for assessment development and use so that inferences based on assumed models can be

supported. The focus of this work is on mechanisms for addressing data-model fit in terms of the adequacy of local independence assumptions in a Bayesian framework.

A number of methods for investigating dimensionality seek to estimate the number of dimensions and determine which items reflect which dimensions. Examples include principle components analysis and similar factor analytic techniques as well as nonparametric approaches to IRT (Zhang & Stout, 1999b). These exploratory techniques are well-suited to situations in which less is known about the substantive theory, observable data, and the interplay between them, such as in the early stages of model development. The approach taken in this work is more confirmatory in the sense that we treat the situation in which the analyst has a model in place (potentially informed by the results of earlier exploratory analyses as well as substantive theory) and wishes to investigate its (in)adequacy in terms of the specified dimensionality.

The balance of this introductory chapter is organized as follows. In the following section we present characterizations of the models studied in the current work. Next, Bayesian modeling and model checking strategies are discussed. Two studies are then introduced.

IRT MODELS FOR DICHOTOMOUS OBSERVABLES

Without loss of generality, the two levels of a dichotomously-scored variable can be coded as 1 (for a correct item response) and 0 (for an incorrect item response). In IRT, each response variable X_{ij} is modeled as a Bernoulli variable, the argument of which is a function of the variables associated with the subject and the observable that defines the probability of a correct response. That is,

$$X_{ij} \sim \text{Bernoulli}\left(P\left(X_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{\omega}_j\right)\right).$$

$P\left(X_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{\omega}_j\right)$ is termed the item response function (IRF). Following common parlance, $\boldsymbol{\omega}_j$ will also be referred to as item parameters. Note that the probability of a correct response need only be expressed conditional only on the examinee's latent variable ($\boldsymbol{\theta}_i$) and parameters for the item in question ($\boldsymbol{\omega}_j$), reflecting the conditional independence assumptions in Equation (1).

Unidimensional Item Response Models

The two parameter logistic model (2-PL; Birnbaum, 1968; see also Hambleton & Swaminathan, 1985) expresses the probability of a subject responding to an item correctly as a function of a single latent variable for the subject, $\boldsymbol{\theta}_i = \theta_i$, and two (item) parameters, b_j and a_j, defining the conditional distribution of X_{ij}. More formally, $\boldsymbol{\omega}_j = \left(b_j, a_j\right)$ and the probability that subject *i* responds to item j correctly is defined as

$$P\left(X_{ij}=1|\theta_i,\boldsymbol{\omega}_j\right) = P\left(X_{ij}=1|\theta_i,b_j,a_j\right) = \frac{\exp(a_j(\theta_i-b_j))}{1+\exp(a_j(\theta_i-b_j))}, \qquad (2)$$

where b_j and a_j are parameters defining the difficulty, and discrimination, respectively, of the IRF.

Multidimensional Item Response Models

Multidimensional IRT (MIRT) models structure the observable responses as dependent on multiple latent dimensions. Though a number of MIRT models have been developed (see van der Linden & Hambleton, 1997 for examples), attention here is restricted to the two models that will be the focus of the first study.

Compensatory MIRT

A compensatory MIRT model for dichotomous items generalizes the model in Equation (2) and specifies the probability of a correct response from examinee *i* responding to item j as

$$P(X_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{\omega}_j)=P(X_{ij}=1|\boldsymbol{\theta}_i,\mathbf{a}_j,d_j)$$
$$=\frac{\exp(a_{j1}\theta_{i1}+a_{j2}\theta_{i2}+\cdots+a_{jM}\theta_{iM}+d_j)}{1+\exp(a_{j1}\theta_{i1}+a_{j2}\theta_{i2}+\cdots+a_{jM}\theta_{iM}+d_j)}, \quad (3)$$

where $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\ldots,\theta_{iM})'$ is a vector of *M* variables that characterize examinee *i* and $\boldsymbol{\omega}_j=(d_j,\mathbf{a}_j)$, where $\mathbf{a}_j=(a_{j1},a_{j2},\ldots,a_{jM})'$ is a vector of *M* coefficients for item j which capture the discriminating power of the associated examinee variables and d_j is an intercept related to the difficulty of the item (Reckase, 1985, 1997a, 1997b; Reckase & McKinley, 1991; see also Ackerman, 1994, 1996).

It is easily seen that the unidimensional model in Equation (2) is just a special case of that in Equation (3). Without loss of generality, assume the single dimension is the first dimension in the MIRT model. A unidimensional model obtains when the $a_{j1}\neq 0$ and, for $m>1$, $a_{jm}=0$. In this case $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$ effectively reduce to the scalars θ_i and a_j, respectively, and $d_j/a_j=-b_j$.

Conjunctive MIRT

A conjunctive MIRT model for dichotomous items combines the latent dimensions in a different manner and specifies the probability of a correct response from examinee i responding to item j as

$$P(X_{ij}=1|\boldsymbol{\theta}_i,\boldsymbol{\omega}_j)=P(X_{ij}=1|\boldsymbol{\theta}_i,\mathbf{b}_j)$$
$$=\prod_{m=1}^{M}\frac{\exp(\theta_{im}-b_{jm})}{1+\exp(\theta_{im}-b_{jm})} \quad (4)$$

where $\boldsymbol{\theta}_i=(\theta_{i1},\theta_{i2},\ldots,\theta_{iM})'$ is a vector of M variables that characterize examinee i and $\boldsymbol{\omega}_j=\mathbf{b}_j=(b_{j1},b_{j2},\ldots,b_{jM})'$ is a vector of M parameters for item j corresponding to the difficulties along the dimensions (Ackerman, 1994; Bolt & Lall, 2003; Embretson, 1984, 1997; Whitely, 1980).

A unidimensional model may also be obtained from the conjunctive MIRT model if the location parameters along all but one dimension decrease without limit. Without loss of generality, assume the single dimension is the first dimension in the MIRT model. Assuming the latent variables are located such that all $\theta_{im}>>-\infty$, a unidimensional model is obtained when the $b_{j1}\neq-\infty$, and for $m>1,\ b_{jm}=-\infty$. Substantively, if the difficulties along dimensions 2,…,M are low enough (i.e., $-\infty$), the probability of a correct response effectively becomes a function of the examinee's latent proficiency and the item's difficulty along the first dimension.

BAYESIAN PSYCHOMETRIC MODELING

Drawing from Schum (1987, 1994) and Jaynes (2003) we maintain that probability based reasoning can play a central role in all forms of inference, including inference in educational measurement and related fields (Mislevy, 1994). Beliefs and uncertainty regarding variables are captured by probability distributions. An inferential process is thus the characterization and evaluation of probability distributions in light of evidence. Once some evidence is observed, Bayesian inference is a framework for the

incorporation and propagation of evidence to arrive at the posterior distribution for unknown variables.

Let $\boldsymbol{\theta}$, $\boldsymbol{\omega}$, and $\mathbf{X}$ denote the complete collections of subject variables, parameters characterizing the psychometric properties of the observables, and the observables themselves, respectively. Taking advantage of the conditional independence assumptions from theory or design, the posterior distribution for $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ given $\mathbf{X}$ may therefore be represented as

$$P(\boldsymbol{\theta},\boldsymbol{\omega}\mid\mathbf{X}) = \frac{P(\boldsymbol{\theta})\times P(\boldsymbol{\omega})\times P(\mathbf{X}\mid\boldsymbol{\theta},\boldsymbol{\omega})}{\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\omega}} P(\boldsymbol{\theta})\times P(\boldsymbol{\omega})\times P(\mathbf{X}\mid\boldsymbol{\theta},\boldsymbol{\omega})d\boldsymbol{\theta}d\boldsymbol{\omega}}$$
$$\propto \prod_i P(\boldsymbol{\theta}_i\mid\boldsymbol{\lambda})\prod_j P(\boldsymbol{\omega}_j\mid\boldsymbol{\eta})\prod_i\prod_j P(X_{ij}\mid\boldsymbol{\theta}_i,\boldsymbol{\omega}_j) \quad (5)$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ are parameters that govern the prior distributions of values for the subject variables and parameters for the observables, respectively.

BAYESIAN NETWORKS

When the number of variables in a problem increases, the application of Bayes' theorem in its textbook form becomes computationally intractable, as the calculation of $P(\mathbf{X})$ (the denominator in Equation (5)) becomes cumbersome. This computation is made easier by the use of discrete variables, in which case the integrals are replaced with summations. However, many psychometric models hypothesize continuous latent variables for examinees (e.g., IRT). In addition, even when all examinee variables are discrete, the parameters characterizing the psychometric properties may be continuous (e.g., Levy & Mislevy, 2004).

More efficient techniques to represent variables and apply Bayes' theorem across a large system of variables have been developed in the form of Bayesian networks (BNs; Brooks, 1998; Jensen, 1996, 2001; Lauritzen & Spiegelhalter, 1988; Pearl, 1988; Spiegelhalter, Dawid, Lauritzen, & Cowell, 1993), which support probability-based reasoning as a means of transmitting complex observational evidence throughout a network of interrelated variables.

A BN is a graphical model of a joint probability distribution over a set of random variables, and consists of the following elements (Jensen, 1996):

- A set of variables (represented by ellipses or boxes and referred to as *nodes*) with a set of *directed edges* (represented by arrows) between nodes indicating the statistical dependence between variables. Nodes at the source of a directed edge are referred to as *parents* of nodes at the destination of the directed edge, their *children*.
- Each variable has a set of exhaustive and mutually exclusive states.
- The variables and the directed edges together form an acyclic directed graph (Brooks, 1998; Jensen, 1996, 2001; Pearl, 1988). These graphs are directed in that the edges follow a "flow" of dependence in a single direction (i.e., the arrows are always unidirectional rather than bi-directional). The graphs are acyclic in that following the directional flow of directed edges from any node it is impossible to return to the node of origin.
- For each endogenous variable (i.e., one with parents), there is an associated set of conditional probability distributions corresponding to each possible pattern of values of the parents.

- For each exogenous variable (i.e., one without parents), there is an associated unconditional probability table or distribution.

The structure of the graph conveys the dependence and conditional independence relationships in the model (Pearl, 1988). Graphical models also facilitate the modular construction of complex models (Almond & Mislevy 1999; Pearl, 1988; Rupp, 2002). In addition to visually representing the model, the graph structures the computations necessary to propagate observable evidence throughout the model to arrive at the posterior distribution (Jensen, 2001; Pearl, 1988).

Technically, BNs are specified as networks of discrete variables only. Network representations of models with continuous variables belong to a broader class of graphical models (Almond & Mislevy, 1999). Models with continuous variables share the above features save that continuous variables have associated (conditional) probability distributions rather than a finite set of states. The key difference is that models with continuous variables require integration to obtain $P(\mathbf{X})$ whereas BNs are less computationally cumbersome, requiring summation over the discrete states. We do not pursue this distinction further and continue to refer to Bayesian models which afford graphical representations and analyses as BNs.

POSTERIOR PREDICTIVE MODEL CHECKING

As noted by several authors (e.g., Sinharay & Johnson, 2003; Williamson, Mislevy, & Almond, 2001), model diagnostics and model criticism remain relatively unexplored aspects of Bayesian psychometric modeling. This study pursues an emerging

and promising set of techniques involving posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Meng, 1994; Rubin, 1984).

The logic of PPMC starts by recognizing that a model has implications for how observable data ought to behave and that data-model fit assessment may be conducted by assessing the features of the observed data in light of the model's implications. Upon estimating a model, a solution is obtained to support inferences. This solution also implies how data ought to behave. The model's solution in a Bayesian framework is the posterior distribution. PPMC proceeds by (a) employing the posterior distribution to empirically characterize the implications for data and then (b) assessing the extent to which the observed data are consistent with those implications. To the extent that the observed data are consistent with the implications, there is evidence of data-model fit. Inconsistencies between the observed data and the implications constitute evidence of data-model misfit. Obtaining the solution (the posterior distribution) has been discussed above. The remaining steps of arriving at the model's implications and then assessing the degree to which the observed data are (in)consistent with the implications are discussed in turn.

The Posterior Predictive Distribution

PPMC analyzes characteristics of the data and/or the discrepancy between the observed data and the model by referring to the posterior predictive distribution. Let $\boldsymbol{\Omega} = (\boldsymbol{\theta}, \boldsymbol{\omega})$ be the full collection of model parameters (i.e., in IRT, the subject variables and parameters characterizing the observables) and recall $\mathbf{X}$ is the full collection of observed data. The posterior predictive distribution is then (Gelman et al., 1996)

$$P\left(\mathbf{X}^{rep} \mid \mathbf{X}\right) = \int_{\mathbf{\Omega}} P\left(\mathbf{X}^{rep} \mid \mathbf{X}, \mathbf{\Omega}\right) P\left(\mathbf{\Omega} \mid \mathbf{X}\right) d\mathbf{\Omega} = \int_{\mathbf{\Omega}} P\left(\mathbf{X}^{rep} \mid \mathbf{\Omega}\right) P\left(\mathbf{\Omega} \mid \mathbf{X}\right) d\mathbf{\Omega},$$

where $P\left(\mathbf{\Omega} \mid \mathbf{X}\right)$ is the posterior distribution (Equation (5)) and $\mathbf{X}^{rep}$ is a set of *replicated data* containing potential but unobserved data values. $\mathbf{X}^{rep}$ are data generated by the model, and may be thought of as data that, though not observed, *could have* come from the model. Uncertainty in the values of the model parameters is contained in the posterior distribution $P\left(\mathbf{\Omega} \mid \mathbf{X}\right)$. By integrating over this posterior, the posterior predictive distribution reflects the averaging over the uncertainty in the model parameters.

Discrepancy Measures

PPMC then comes to characterizing and assessing the observed data in relation to the posterior predictive distribution. In performing PPMC, a set of functions called *discrepancy measures* are defined to capture relevant features of the data and/or the discrepancy between data and the model. We denote discrepancy measures as $D\left(*, \mathbf{\Omega}\right)$, where the arguments are a data set (observed or replicated) and the model parameters.[1] The discrepancy measures should be chosen to reflect important features of the model and the inferences in question. Meaningful differences between the *realized discrepancies* $D\left(\mathbf{X}, \mathbf{\Omega}\right)$, based on the observed data, and the distribution of $D\left(\mathbf{X}^{rep}, \mathbf{\Omega}\right)$, based on the posterior predictive distribution, are indicative of data-model misfit.

[1] It is possible to distinguish between functions that depend only on the data and those that depend on both the data and the model parameters. When this distinction is made, the former are referred to as *test statistics*. This distinction is not pursued here; we refer to all quantities used in PPMC as discrepancy measures.

Markov Chain Monte Carlo Estimation

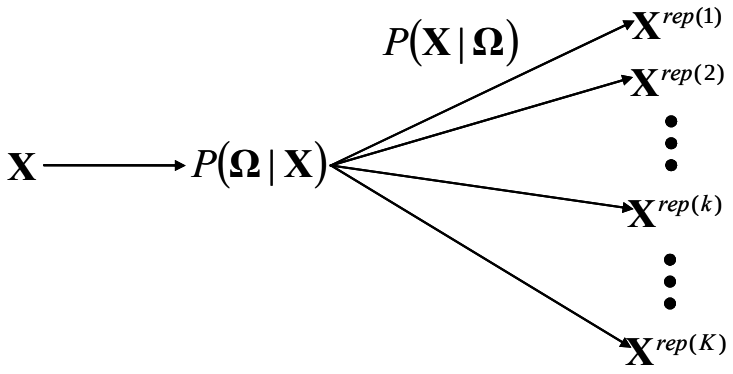
Analytical procedures for evaluating posterior distributions and posterior predictive distributions are available for simple problems but are computationally intractable for complex, multivariate problems often encountered in educational and psychological measurement. Taking advantage of the proportionality relation in Equation (5), Markov chain Monte Carlo (MCMC; Gelfand & Smith, 1990, Gilks, Richardson, & Spiegelhalter, 1996b; Smith & Roberts, 1993; Tierney, 1994) estimation consists of drawing possibly dependent samples from the distribution of interest and as such provides an appropriate framework for computation in Bayesian analyses (Brooks, 1998; Gelman, Carlin, Stern, & Rubin, 1995). A complete treatment and description of MCMC estimation is beyond the scope and intent of this work. Specifics regarding the MCMC estimation techniques employed in these studies are given in subsequent sections.

For the current purpose as it relates to PPMC, it is important to recognize that MCMC estimation consists of drawing values for variables from a series of distributions that is in the limit equivalent to drawing values from the true posterior distribution (Gilks, Richardson, & Spiegelhalter, 1996a; see Roberts, 1996 for regularity conditions). To empirically sample from the posterior distribution, it is sufficient to construct a Markov chain that has the posterior distribution as its stationary distribution. An arbitrarily large number of iterations may be performed resulting in simulated values of the unknown parameters that form an empirical approximation to the posterior distribution.

MCMC estimation is also well suited for compilation of the posterior predictive distribution. In each replication of the MCMC process, values are drawn for the model parameters $\boldsymbol{\Omega}$. The lone additional step necessary to arrive at the posterior predictive

distribution is to generate a new data set via the model. Figure 3 depicts the structure of this process. In conjunction with the prior distribution for the unknown parameters $\mathbf{\Omega}$ and the conditional distribution of the data given the unknown parameters (not pictured), observing the data $\mathbf{X}$ results in the posterior distribution $P(\mathbf{\Omega} \mid \mathbf{X})$. For each of *K* draws from the posterior of $\mathbf{\Omega}$ resulting in $\{\mathbf{\Omega}^1, \ldots, \mathbf{\Omega}^K\}$, generate $\{\mathbf{X}^{rep(1)}, \ldots, \mathbf{X}^{rep(K)}\}$ accordingly from the model, $\mathbf{X}^{rep(k)} \sim P(\mathbf{X} \mid \mathbf{\Omega} = \mathbf{\Omega}^k)$. The set of generated $\mathbf{X}^{rep}$ constitutes an empirical approximation to the posterior predictive distribution.

Figure 3: Structure of the process to obtain the posterior predictive distribution.



We are left with the observed data, $\mathbf{X}$, and the set of replicated data sets, $\mathbf{X}^{rep(1)}, \ldots, \mathbf{X}^{rep(K)}$. The operative question is then whether the features in the observed data, for which the adequacy of the model is still an open question, look like the patterns of those features in the replicated data, for which we know the model is exactly right by construction. If the features of the observed data are consistent with the patterns in the replicated data, there is evidence of data-model fit. Departures from the patterns in the replicated data are evidence of data-model misfit.

To represent the patterns in the posterior predictive distribution, the discrepancy measure is computed for each draw from the posterior and corresponding replicated data set. The set of values obtained from evaluating the discrepancy measure in the posterior predicted data sets, $\left\{D\left(\mathbf{X}^{rep1}, \mathbf{\Omega}^{1}\right), \ldots, D\left(\mathbf{X}^{repK}, \mathbf{\Omega}^{K}\right)\right\}$, constitutes an empirical approximation to the posterior predictive distribution of $D(*, \mathbf{\Omega})$.

Posterior Predictive P-Values

A useful mechanism for summarizing information in PPMC is a posterior predictive p-value (PPP-value; Gelman et al., 1996; Meng, 1994). For each discrepancy measure, the distribution of the values of the discrepancy measure based on the posterior predictive distribution represents a null distribution against which the values obtained from the observed data, the *realized* discrepancies, may be compared. The PPP-value is the tail-area of the posterior predictive distribution of the discrepancy measure corresponding to the observed value for the discrepancy measure:

$$\begin{aligned} \text{PPP - value} &= P\left(D\left(\mathbf{X}^{rep}, \mathbf{\Omega}\right) \geq D\left(\mathbf{X}, \mathbf{\Omega}\right) \mid \mathbf{X}, \mathbf{H}\right) \\ &= \int I\left[D\left(\mathbf{X}^{rep}, \mathbf{\Omega}\right) \geq D\left(\mathbf{X}, \mathbf{\Omega}\right)\right] P\left(\mathbf{X}^{rep} \mid \mathbf{\Omega}\right) P\left(\mathbf{\Omega} \mid \mathbf{X}\right) d\mathbf{X}^{rep}\, d\mathbf{\Omega} \\ &\approx \frac{\sum_{k=1}^{K} I\left[D\left(\mathbf{X}^{rep(k)}, \mathbf{\Omega}^{k}\right) \geq D\left(\mathbf{X}, \mathbf{\Omega}^{k}\right)\right]}{K} \end{aligned} \tag{6}$$

where **H** is the (null) hypothesis that the model holds in the population and $I[*]$ is the indicator function that takes on a value of one when its argument is true and zero otherwise. The conditioning on **H** makes explicit the role of the model in defining the posterior distribution of $\mathbf{\Omega}$ and the posterior predictive distribution of $\mathbf{X}^{rep}$.

Substantively, the PPP-value may be thought of as the probability of obtaining a value for the discrepancy measure in the posterior predictive data (i.e., given that the model holds) that is larger than the observed value of discrepancy measure. PPP-values near .5 indicate that the realized discrepancies fall in the middle of the distribution of discrepancy measures based on the posterior predictive distribution and are indicative of data-model fit. PPP-values near 0 or 1 suggest that the observed data are inconsistent with the posterior predictive distribution and hence are indicative of data-model misfit. More specifically, PPP-values near 0 indicate that the realized values are far out in the upper tail of the distribution, which indicate that the model is *underpredicting* the quantity of interest. By the same logic, PPP-values near 0 indicate that the model is *overpredicting* the quantity of interest.

As the last expression in Equation (6) indicates, in a simulation environment such as MCMC, the PPP-value may be approximated by the proportion of the K draws in which the predicted discrepancy $D\left(\mathbf{X}^{rep(k)}, \mathbf{\Omega}^{k}\right)$ exceeds the realized discrepancies $D\left(\mathbf{X}, \mathbf{\Omega}^{k}\right)$. The empirical approximation improves as K increases.

Discussion

PPMC is a powerful and flexible tool for model criticism and holds many advantages over traditional techniques. In frequentist approaches, considerable work may be needed to derive sampling distributions, which may not be well-defined. As reviewed in greater detail later, there is ambiguity regarding the sampling distributions of some of the more popular indexes used in assessing local dependence (Chen & Thissen 1997; Yen, 1993).

Furthermore, frequentist null sampling distributions for most discrepancy measures are justified only asymptotically, as they depend on unknown parameters for which consistent estimates (e.g., maximum likelihood estimates) are substituted (Meng, 1994). PPMC makes no such appeal to asymptotics and does not require other regularity conditions associated with frequentist model checking (e.g., non-zero cell counts, Fu, Bolt, & Li, 2005; Sinharay, in pressa).

Because there is no appeal to asymptotic behavior of well behaved functions, there is no need to restrict attention to measures for which sampling distributions can be determined (Jannsen, Tuerlinckx, Meulders, & De Boeck, 2000). The values of discrepancy measures based on the posterior predicted data represent an empirically constructed reference distribution. As a consequence, the analyst is free to choose from a broad class of functions, including those that pose difficulties for traditional model checking techniques (see Sinharay & Stern, 2003, for an example of one such function).

A final theme of PPMC is the modeling of uncertainty. The use of point estimates for parameters in model criticism ignores (that is, understates) the uncertainty in the sampling distributions of discrepancy measures (Meng, 1994). Recent recognition and appreciation of this has led to the desire to incorporate the uncertainty in parameter estimates in a frequentist approach to item fit (Donoghue & Hombo, 1999). The solution to this issue has proven to be considerably difficult to obtain, and is localized to the specifics of the problem at hand (Donoghue & Hombo, 1999, 2001, 2003). PPMC automatically incorporates the uncertainty by using the posterior distribution, rather than point estimates, of the model parameters and may easily be instantiated in a variety of settings.

A closely related frequentist strategy that circumvents the need to derive sampling distributions involves bootstrapping (Mooney & Duval, 1993). Bootstrapping is a process by which the observed data are re-sampled (with replacement) to create simulated samples. In essence, the observed data are treated like a population and samples are repeatedly drawn. A discrepancy measure is calculated for the observed data and then calculated for each of the bootstrapped samples. The distribution of the discrepancy measure in the bootstrapped samples forms a reference distribution against which the value based on the observed sample may be compared. A p-value is obtained as the proportion of times the value of the discrepancy measure in the bootstrapped data sets is more extreme than the value of the discrepancy measure based on the observed data set.

A key development in this line of model checking came in the form of the parametric bootstrap (Beran & Srivastava, 1985; Bollen & Stine, 1993). Straightforward re-sampling the observed data set need not result in the desired sampling distribution, as there is no guarantee that the observed data (which are essentially treated as a population) represent a null population (Bollen & Stine, 1993).[2] In the parametric bootstrap, the observed data are transformed to perfectly fit the model; sampling from this transformed data approximates sampling from a population that is consistent with the model (Beran & Srivastava, 1985; Bollen & Stine, 1993).

Transforming the observed data before re-sampling is a data-based approach to conducting a parametric bootstrap. A model-based strategy leaves the observed data untransformed and obtains samples by generating data directly from the model, using parameter estimates as the values of the parameters. Data sets are generated from the

[2] Indeed, to assume that the observed data are consistent with the model in justifying the bootstrapped reference distribution would beg the question of data-model fit.

model and the discrepancy measure is calculated for each. The distribution of these values constitutes a reference distribution against which the observed value may be compared.

PPMC has much in common with the parametric bootstrap, in particular, the model-based parametric bootstrap. Both techniques construct reference distributions empirically using data sets that are generated from the solution to the model. The key difference lies in what each technique considers the "solution to the model" to be. In the parametric bootstrap, the solution is the set of point estimates for the model parameters. As argued above, an approach that treats the model parameter estimates as known values ignores the uncertainty in the model parameters, and is justified asymptotically if the point estimates are consistent (e.g., maximum likelihood estimates). In contrast, the solution in a Bayesian analysis is the full posterior distribution. Data are generated not from one set of values for the model parameters (i.e., point estimates) but rather by taking draws from the full posterior distribution of the model parameters. Thus PPMC incorporates the uncertainty in the model parameters. As a consequence, the PPP-value may be thought of as the integration of the frequentist p-value over the posterior distribution of (uncertainty in) the model parameters (Meng, 1994).

Although the advantages of PPMC pertain to psychometric models generally, they are particularly relevant to innovative or cognitively-based models which posit atypical relations of complex form, such as inhibition effects explained in Chapter 5, ceiling effects, leaky conjunctive effects and other complex structures (Almond et al., 2001; Arminger & Muthén, 1998; Levy & Mislevy, 2004; Mislevy et al., 2002). A Bayesian framework offers flexible approaches to modeling the assessment, structuring the

statistical model, and estimation (Mislevy & Levy, in press). Likewise, PPMC offers flexible approaches to data-model fit assessment and model criticism that are suitable for models with atypical relations among the variables.

STUDY PURPOSES

The current work investigates the efficacy of PPMC for conducting model criticism in light of the presence of inadequately modeled multidimensionality. The first study pursues this in the context of popular IRT models. As such, the results have implications for dimensionality assessment and model criticism for models frequently employed in operational assessments. The second study is based on the multidimensional statistical model for a complex, simulation-based diagnostic assessment (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). Thus in addition to implications for popular assessment situations, the results of this work speak to applications of atypical models in innovative contexts.

Owing to the communalities among seemingly different psychometric and statistical models in terms of dimensionality and conditional independence assumptions (Mellenbergh, 1994; Rupp, 2002), the results of this work have implications for seemingly different modeling frameworks in addition to the IRT models studied here, including factor analysis, SEM, and LCA. More specifically, the results of this work ought to be suggestive of model criticism in such other modeling frameworks employed in educational measurement and other social sciences.

A dominant theme in PPMC is that of flexibility, both in terms of the discrepancy measures chosen and the types of models that can be subjected to PPMC as a form of model criticism. For the methodologist, PPMC represents an emerging set of tools to be

studied in alternative contexts (some of which have been alluded to above). For the applied analyst, PPMC represents a model checking strategy that may be applied to a variety of model types. Rich substantive theories often include features (e.g., relations among variables, complex constraints) that do not conform to traditional modeling paradigms. Explicitly building models in accordance with substantive theory often leads to the emergence of complex models with intricate hypotheses, the likes of which are not easily evaluated using traditional model checking tools. PPMC represents a flexible set of model checking tools with the potential to go beyond the limits of traditional statistical requirements for data-model fit assessment and help bridge the divide between statistical modeling and substantive theory regarding the phenomena of interest.

CURRENT STUDIES

The goal of the current work is to investigate the potential for applying PPMC to perform model criticism in the presence of inadequately modeled multidimensionality. The aims are to illustrate and compare the performance of several functions used in conducting PPMC under several conditions. The current work consists of two studies, introduced briefly here in turn.

Study 1: Posterior Predictive Model Checking for Multidimensionality in IRT

The first study examines the use of several functions for diagnosing local dependence among unidimensional IRT models for dichotomous observables due to multidimensionality. As discussed above, many psychometric phenomena can be framed as multidimensionality. Although PPMC has seen use in psychometric models (e.g., Fu et al., 2005; Hoijtink, 1998, 2001; Scheines, Hoijtink, & Boomsma, 1999; Sinharay,

2005, in pressa, in pressb; Sinharay & Stern, 2003) and has attracted methodological attention (e.g., Sinharay, in pressa; Sinharay, Johnson, & Stern, in press), no study has explored PPMC for dimensionality assessment in the context of systematically manipulating factors that influence dimensionality.

To that end, a Monte Carlo study is conducted in which a number of key factors relevant to dimensionality are varied, and multiple data sets are generated under each condition to facilitate an examination of the application of PPMC to situations in which a unidimensional model is hypothesized but the data exhibit multidimensionality. Full details of the study are given in Chapter 4. The results of this study have immediate implications for common IRT models in use today as well as for emerging innovative models, such as those discussed in the second study. Furthermore, owing to the communalities of data-model fit assessment across different types of psychometric models (McDonald & Mok, 1995) the results of this study should be suggestive for techniques for addressing questions of unaccounted for dependence among observations in other psychometric models (e.g., SEM, LCA).

Study 2: Posterior Predictive Model Checking for Multidimensionality in BNs

The second study builds upon the first study by applying PPMC to investigate complex multivariate associations in a BN. The models studied here are more complex than those in study 1. More specifically, both the data-generating model and the estimated model are multidimensional. Though a number of techniques exist for assessing assumptions of dimensionality (e.g., Nandakumar & Stout, 1993; Stout 1987), many are limited in that they are applicable only to checking models assumed to be unidimensional (Swaminathan, Hambleton, & Rogers, in press). It is maintained that the

PPMC techniques conducted in these studies are applicable to model criticism of both unidimensional and multidimensional models. This study is more exploratory in nature than the predecessor; more complete details are given in Chapter 5.

CHAPTER 2: FURTHER REVIEW OF EXISTING WORK

This chapter provides further background on key concepts investigated in this work. In particular, the sections in this chapter review (a) MIRT, (b) Bayesian psychometric modeling, with an emphasis on IRT and BNs, and (c) relevant techniques (Bayesian and otherwise) for model criticism for local dependence and dimensionality assessment.

MULTIDIMENSIONAL IRT

Compensatory MIRT models were popularized in their form given here principally by Reckase (1985) in his definition of multidimensional item difficulty. Definitions of discrimination and information in multidimensional space followed shortly thereafter (Reckase & McKinley, 1991; see also Ackerman, 1994, 1996).

Compensatory MIRT models have been applied to understand and model test data in numerous ways. Assessments are often constructed to reflect multiple, possibly related sets of knowledge, skills, and abilities. MIRT models are useful for situations in which multiple latent proficiencies are hypothesized to influence item performance, even when not all of the latent proficiencies are of interest (Ackerman, 1992, 1994, 1996; Bolt & Stout, 1996; Mellenbergh, 1994; Roussos & Stout, 1996; Shealy & Stout, 1996; Stout et al., 1996).

Further, MIRT models are useful in more exploratory settings, where hypotheses regarding the number and nature of the underlying dimensions are absent. Ackerman (1996) described the use of graphical techniques to explicate the behavior of items in a multidimensional setting. Such techniques may be used to (a) examine the number of dimensions underlying a set of items, (b) convey the relationship between an item and the

underlying dimensions, (c) identify the single best dimension that is measured by a set of items, (d) suggest substantive interpretations for latent dimensions, (e) create subtests, and (f) examine differences between multiple forms (Ackerman, 1996). In passing it is noted that the desire to model the (intended) assessment of multiple traits has motivated advances in test assembly (van der Linden, 1996), adaptive testing (Luecht, 1996; Segall, 1996), linking (Davey, Ohima, & Lee, 1996), and task design (Embretson, 1985; Irvine & Kyllonen, 2002).

Models for multiple latent dimensions may be applied to situations in which several item formats (e.g., passage-based, multiple choice, fill in the blank, and constructed response for assessments of verbal proficiency) are employed to assess one or more dimensions of interest. A MIRT model would then include latent variables for the dimensions of interest and latent variables representing the formats. Treating the formats more generally as different methods of assessment, the pattern of free and restricted coefficients would then represent a MIRT representation of a multi-trait multi-method analysis (Campbell & Fiske, 1959). The values of the coefficients would convey the effects of different measurement methods relative to each other and to the dimension(s) of interest. These and similar analyses have been shown to be useful in analyzing assessments containing testlets (Bradlow et al., 1999; Li, Bolt, & Fu, in press).

MIRT models need not be restricted to the case of only latent examinee variables. Zwinderman (1997) applied a MIRT model using observed auxiliary dimensions to investigate the influence of the testing process itself, that is, on the influence of items presented earlier on those presented later, as may be natural in situations in which feedback is presented to the examinee (e.g., Vispoel, 1998).

The literature on conjunctive models is considerably less deep and diverse, though it is growing. Conjunctive MIRT models were first proposed by Sympson (1978) and Whitely (1980) and have seen a number of developments in terms of extensions and connections to cognitive underpinnings of tasks (Embretson, 1984, 1997). Challenges in estimating such models have prevented them from gaining the popularity of compensatory models. Recent advances in MCMC techniques allow for the estimation of complex conjunctive models. Bolt and Lall (2003) demonstrated MCMC estimation for a conjunctive model and provided code for estimating the model in the flexible freeware WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003). Accordingly, it is hypothesized that applications of conjunctive models will become more popular in the immediate future.

Many of the applications of compensatory MIRT models described above are warranted for conjunctive MIRT models. For example, viewing differential item functioning as the presence of an auxiliary dimension is consistent with conjunctive MIRT. Regarding applications that depend on the IRF (e.g., test assembly, linking), developments for conjunctive models will hopefully follow their compensatory counterparts soon.

The key difference between compensatory and conjunctive MIRT lies in their assumptions regarding the underlying processes that give rise to observable performance. In compensatory MIRT models, the contributions of the multiple dimensions are additive. If an examinee lacks a certain skill or level of proficiency, other skills or proficiencies can compensate. In contrast, the lack of a skill or a certain level of proficiency cannot be made up for in conjunctive MIRT models. Rather, it is necessary to have a sufficient

level of each proficiency in order to successfully perform the task. As such, compensatory MIRT models are appropriate for compensatory or disjunctive response processes and conjunctive MIRT models are appropriate for conjunctive response processes (Bolt & Lall, 2003).

In closing this discussion of MIRT models, we note the connection between compensatory MIRT and other psychometric models (Reckase, 1997b). A factor analytic perspective views MIRT as the factor analysis of dichotomous variables (e.g., McDonald, 1997; Mislevy, 1986b; Muthén & Christofferson, 1981; Reckase, 1997b) and uses the normal-ogive model rather than the logistic model. Formal equivalence between the factor analysis of dichotomous variables and normal-ogive IRT models was proved by Takane and DeLeeuw (1987). While the use of the normal-ogive rather than a logistic IRT model affords this connection and may be easier in some instances of MCMC estimation (Albert, 1992), the logistic formulation presented above will be used for three reasons. First, the relative ease of estimation for normal-ogive models only holds under certain assumptions regarding the prior distributions (Maris & Bechger, in preparation) and we do not wish to be restrictive. Second, logistic IRT models are more prevalent in operational assessment. Third, logistic models afford easier interpretations and afford connections to LCA models that employ logistic relations. What remains encouraging, however, is the potential for extensions and applications of this research to factor analytic models both common (compensatory) and uncommon (conjunctive).

BAYESIAN PSYCHOMETRIC MODELING

Bayesian psychometric modeling and estimation strategies have their roots in early applications to classical test theory (Novick, Jackson, & Thayer, 1971) and factor

analysis (Martin & McDonald, 1975). Bayesian modeling and estimation strategies are receiving an increasing amount of attention in IRT (Albert, 1992; Albert & Chib, 1993; Almond & Mislevy, 1999; Béguin & Glas, 2001; Bolt & Lall, 2003; Patz & Junker, 1999a, 1999b; Rupp, Dey, & Zumbo, 2004), LCA (Hoijtink, 1998; Hoijtink & Molenaar, 1997), factor analysis and SEM, (Arminger & Muthén, 1998; Lee, 1981; Lee & Song, 2004; Rowe, 2003; Scheines et al., 1999), multilevel modeling (Fox & Glas, 2001; Seltzer, Wong, & Bryk, 1996), cognitive diagnosis models (Martin & VanLehn, 1995; Mislevy, 1995, Williamson, 2000; Williamson et al., 2001), and the psychometric modeling community generally (Mislevy, 1994).

Bayesian techniques have proven useful in modeling assessment settings such as adaptive testing (Almond & Mislevy, 1999) and accounting for psychometric phenomena such as testlet effects (Bradlow et al., 1999; Li et al., in press). The emergence of research specialized to particular aspects of Bayesian modeling such as investigations in MCMC convergence assessment (Sinharay, 2004) and model fit (e.g., Sinharay, 2005) as applied to psychometric situations constitutes another step in the maturation of Bayesian modeling as an approach to psychometric problems.

Bayesian IRT

Early work on fully Bayesian IRT models was conducted by Swaminathan and Gifford (1982, 1985, 1986). Mislevy (1986a) detailed the principles, properties, and key elements of a fully Bayesian analysis of IRT models. More recent theoretical work has established the conditions of propriety of posterior distributions for IRT models (Ghosh, Ghosh, Chen, & Agresti, 2000). With a firm foundation, research shifted towards model estimation strategies and applications to complex assessment scenarios.

Albert (1992) introduced a data-augmentation Gibbs sampling solution for the two parameter normal-ogive (2-PNO) model; the algorithm was extended to handle polytomous data by Albert and Chib (1993). Sahu (2002) described a similar data-augmented Gibbs sampling approach for the three parameter normal-ogive model. Data-augmentation Gibbs sampling for the 2-PNO model works when normal priors are used for item parameters, yielding tractable full conditional distributions. Under alternate priors, the full conditionals are not tractable (Maris & Bechger, in preparation). Maris and Bechger (in preparation) describe a data-augmentation Gibbs sampler for a transformed parameterization to create tractable full conditionals (that does not depend on the choice of priors) for the 2-PL model and suggest extensions for more complex models.

The key turning point in the use of MCMC estimation for Bayesian IRT models was the work of Patz and Junker (1999a, 1999b). They introduced a Metropolis-Hastings-within-Gibbs approach which was much more flexible than the Gibbs sampling approach of Albert (1992). As a consequence, logistic IRT models for dichotomous data, polytomous data, missing data, and rater effects could be handled (Patz & Junker, 1999a; 1999b) without restrictions on the priors.

The flexibility of MCMC estimation of IRT models has allowed the expansion to more complex models, such as those involving testlets (Bradlow et al., 1999; Li et al., in press), multilevel models (Fox & Glas, 2001), hierarchical models for mastery classification (Jannsen et al., 2000), and multidimensional models both compensatory and conjunctive (Béguin & Glas, 2001; Bolt & Lall, 2003). See Rupp et al., (2004) for a recent review of applications and extensions.

Bayesian Networks

BNs offer flexible modeling opportunities and have seen applications in settings from devising expert systems (Spiegelhalter et al., 1993) to weather forecasting (Edwards, 1998). In the psychometric community, BNs have frequently been applied to cognitive diagnosis models and closely related latent class models in which a finite set of discrete latent attributes are hypothesized to underlie performance (Hoijtink, 1998; Hoijtink & Molenaar, 1997; Mislevy, 1994, 1995; Mislevy & Gitomer, 1996; Mislevy et al., 2002; Martin & VanLehn, 1995; Williamson, 2000; Williamson et al., 2001). The use of BNs to model item responses and discrete latent variables or attributes overlaps with the traditional uses of BNs for modeling discrete latent variables.

More recently, applications have involved modeling continuous latent variables. de la Torre and Douglas (2004) employed a continuous latent variable to model the association among the attributes in a cognitive diagnosis model. Mislevy (1994) and Almond and Mislevy (1999) placed IRT in a BN framework, unifying themes from graphical modeling and IRT. Janssen et al. (2000) likewise treated IRT using graphical network approaches.

The flexibility of BNs to manage multiple, possibly conflicting forms of evidence in the form of large numbers of observables and latent variables makes BNs well-suited to many innovative assessments. Almond and Mislevy (1999) described a graphical modeling approach to adaptive testing. Almond et al. (2001) introduced several complex relationships in terms of a BN. See Mislevy et al. (2002) for further illustrations. BNs have also found a home in assessments for intelligent tutoring systems (Conati, Gertner, VanLehn, & Druzdzel, 1997; Mislevy & Gitomer, 1996).

In concluding this section on Bayesian psychometric modeling, we note that the use of an acyclic graphical model that contains only directed edges conflicts in certain ways with other path diagrammatic modeling approaches. In SEM, for example, bi-directional arrows are common to signal that a dependence between variables is assumed without a hypothesized direction (e.g., Bollen, 1989). A number of options exist for converting graphs with bi-directional connections to one with only directed connections. In the absence of an assumed directional influence (a) a directed edge may be employed provided it does not change the implications of the model (Lee & Hershberger, 1990; Raykov & Penev, 1999; Stelzl, 1986), (b) an additional variable may be introduced (e.g., a higher-order factor; Thurstone, 1947) to account for the association, or perhaps most consistent with the intent of an undirected association, (c) the variables may be specified as following a multivariate distribution.

POSTERIOR PREDICTIVE MODEL CHECKING

Though Bayesian modeling and estimation strategies have reached a state of maturity, what has *not* yet fully matured is model checking in Bayesian psychometric modeling. A fully Bayesian analysis performs model comparisons by the evaluation of Bayes factors, a topic that has received some attention in psychometric models (Raftery, 1993; 1996; Sahu, 2002). The situation here is slightly different. The context of these studies is that, as in an operational assessment, there is a desired model to be employed and the analyst wishes to evaluate the extent to which the model conforms to the data. In passing, we note that recent work has shown PPMC to be effective for model comparison and selection when competing models are considered (Li et al., in press).

Box (1980) suggested the use of the prior predictive distribution (i.e., the marginal distribution of the data) in the calculation of a p-value for model criticism. Two shortcomings of this approach are that such analyses cannot be conducted when the prior distribution is improper and, as is obvious, the choice of a prior has a strong influence, even for large data sets (Gelman et al., 1996). As a consequence, more attention has focused on evaluating discrepancy measures with respect to the posterior distribution.

Gelman et al. (1996) and Sinharay and Johnson (2003) traced PPMC to Guttman (1967), though a modern Bayesian description is given by Rubin (1984), who used the posterior predictive distribution of a statistic as a reference distribution against which to compare the observed value, resulting in the tail-area probability (PPP-value). In discussing model checking more broadly, Meng (1994) noted that unless a test statistic is a pivotal quantity, the classical definition of a p-value cannot be calculated, as the sampling distribution depends on unknown parameters. The typical solution of inserting a maximum likelihood point estimate is only justified asymptotically, and relies on consistency results from maximum likelihood theory (e.g. White, 1982). In this sense, many statistics used in criticizing models may be thought of as discrepancy measures that evaluate the discrepancy between the observed data and the "best fitting" population parameters (Meng, 1994). Though this approach handles the functional dependence of the statistic on unknown parameters, it ignores the dependence of the sampling distribution of the test statistic on the unknown parameters; the PPP-value (Equation (6)) accommodates this dependence (Meng, 1994). Indeed, just as the posterior predictive distribution averages over the uncertainty in the model parameters, the PPP-value is an

average p-value over the posterior distribution of the unknown parameters (Gelman et al., 1996; Meng, 1994).

However, PPP-values are not free from criticism. Robins, van der Vaart, and Ventura (2000) showed that PPP-values are not uniformly distributed under null conditions, even asymptotically. Rather, the distribution is centered at .5 but less dispersed than a uniform distribution (Meng, 1994; Robins et al., 2000; Rubin, 1996). The result is that employing PPP-values to conduct hypothesis testing leads to conservative inferences. In a hypothesis testing framework, a (two-tailed) test with significance level α is performed by rejecting the null hypothesis of data-model fit if the PPP-value is less than $\alpha/2$ or is greater than $1-\alpha/2$. Owing to the conservativeness of PPP-values, the actual Type I error rate for such a test will be less than α. In evaluating IRT models, Sinharay et al. (in press) found PPP-values to be below nominal values. Similar results were found by Fu et al. (2005) in the context of cognitive diagnosis models.

This conservativeness is a result of the double use of the data for estimation and model checking (Bayarri & Berger, 2000; Berkhof, van Mechelen, & Gelman, 2004; Draper, 1996; Robins et al., 2000). In calibrating the model and conducting PPMC, the model is trained to the data at hand and then checked by the data. In other words, use of the data to estimate and then test the model gives the model the best chance at seeming appropriate.

Bayarri and Berger (2000) proposed alternative model checking techniques, but they are limited in that they may be more difficult to perform and interpret or may be restrictive in the types of discrepancy measures they support (Berkhof et al., 2004).

The orientation of this work is therefore to treat PPMC and PPP-values as pieces of statistical evidence for, rather than a test of, data-model (mis)fit (Berkhof et al., 2004; Gelman et al., 1996; Stern, 2000). It is maintained that rejecting a model based on a PPP-value beyond some cutoff is not warranted, regardless of whether the PPP-value is uniformly distributed under the null. This is especially true when the model under consideration is theoretically justified and/or the result of earlier model criticism and refinement in more exploratory settings.

Model criticism is a complicated process that depends on many criteria. PPMC and PPP-values are better 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). One advantage of PPMC is that any function of interest may be investigated. Functions should be chosen to reflect the (possibly multiple) feature(s) of interest; concluding that a model adequately captures some but not all features of the data is not uncommon. Accordingly, an approach that considers numerical PPP-values in conjunction with other model checking techniques (including other statistical techniques, e.g., graphical procedures, Gelman et al., 1996; Jannsen et al., 2000; Sinharay, 2005) is necessary for model criticism. Such an interpretation is consistent with the approach to model criticism that views statistical diagnostics as a component of a larger enterprise, guided by substantive theory, aimed at evaluating model adequacy (Sinharay, 2005).

Note that this attitude is shared by analysts (operating outside of the Bayesian framework) for detecting and diagnosing violations of local independence. The current work follows Chen and Thissen (1997) in maintaining that “Any meaningful interpretation of the LD [local dependence] indexes requires skill and experience in IRT

analysis and close examination of the item content" (p. 288) and likewise follows Zenisky et al. (2003) in maintaining that "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" (pp. 17-18).

Even from a perspective that views PPMC and PPP-values as diagnostic measures, uniformity of PPP-values under null conditions may be desirable (Berkhof et al., 2004). The purpose of this investigation is the extent to which PPMC can be employed to diagnose model failure due to failure to adequately model multidimensionality. To this end, the behavior of PPP-values under null and non-null conditions will be examined. As discussed below, several studies have successfully employed PPMC and PPP-values to critique models.

LOCAL DEPENDENCE AND DIMENSIONALITY ASSESSMENT

In this section, popular methods for assessing violations of local independence are briefly reviewed. Yen advocated the use of Q_3 (Yen, 1984; 1993), a measure of the correlation of residual difference between observed and expected responses for an item, and further provided motivation for its expectation under null conditions based on normality of the sampling distribution. Zenisky et al. (2003) also found Q_3 to be a useful index. Chen and Thissen (1997) also found Q_3 to be an effective index, though they demonstrated the normality assumption is questionable.

Likewise, Chen and Thissen (1997) found X^2 and G^2 measures for item-pairs to be comparable but question the appropriate sampling distribution for these measures as well. In 2x2 tables (as arise from the cross-tabulation of two dichotomous items), X^2 and G^2

tests for independence should have one degree of freedom. However, the inclusion of discrimination parameters results in what "may be described as the loss of a fraction of the one degree of freedom for the test of independence" (Chen & Thissen, 1997, p. 269). Accordingly they found null distributions to have empirical means less than one.

In sum, the assumed null distributions for popular measures of item dependence are not only asymptotically justified, but at best are only approximations. However, PPMC is not subject to these difficulties. There is no need to work out the sampling distribution (asymptotic or otherwise) of the chosen discrepancy measure. Rather, the reference distribution is constructed from the posterior predictive distribution. In lamenting that null distributions were not known, Chen and Thissen (1997) employed empirical distributions from locally independent data. This is in the same spirit as PPMC, in which the model, which assumes local independence, generates the posterior predicted data sets.

A number of tools based on conditional covariance theory (Bolt, 2001; Nandakumar & Stout, 1993; Stout 1987; Stout et al., 1996; Zhang & Stout, 1999a, 1999b) have been developed to perform dimensionality assessment and estimation. The theoretical framework of conditional covariance theory and its application to studying multidimensionality is reviewed in greater detail in Chapter 3. Two popular tools based on conditional covariance theory are the DIMTEST and DETECT procedures. The DIMTEST procedure (Nandakumar & Stout, 1993; Stout 1987) evaluates the conditional covariance between judiciously selected subsets of a test to assess departures from unidimensionality. The DETECT procedure (Zhang & Stout, 1999b) seeks to partition

the items into clusters such that approximate simple structure obtains within each partition.

DETECT is more exploratory in nature; DIMTEST is an explicit test of essential unidimensionality (Stout, 1987) and is more in line with the confirmatory settings of interest in the current work. DIMTEST is limited in that (a) the analyst must identify a subtest for directed investigation of departure from unidimensionality, (b) the test must be long enough to support the division of the test into subtests, and (c) it is designed to test a hypothesis of unidimensionality (Hoijtink, 2001; Swaminathan et al., in press). This last point is more salient in light of desires for assessments that target complex domains with interrelated proficiencies (e.g., Behrens et al., 2004).

PPMC FOR MODEL FIT AND DIMENSIONALITY ASSESSMENT

Turning to applications of PPMC for model criticism, Fu et al. (2005) used PPMC to evaluate item fit for cognitive diagnosis models. They examined X^2 and G^2 measures for item-pairs as well as analogous measures for items individually and items with total scores. Additionally, they considered the bivariate item covariance discrepancy argued for by McDonald and Mok (1995) at both the item-pair level and aggregated to the model level. Relevant to this study, Fu et al. (2005) found that bivariate measures are more successful than univariate measures, and further found the bivariate item covariance discrepancy to perform the best. Consistent with theoretical results (Robins et al., 2000), they found the empirical Type I error rates to be less than nominal levels.

Li, et al. (in press) applied PPMC techniques to criticize models on the basis of the influence of testlets and employed X^2 measures at the item, testlet, and test level, as well as the odds ratio for item-pairs. They concluded that the X^2 measures were not

useful in determining the correct model, but that the odds ratios were very effective. Interestingly, they also found that, in the presence of multiple models under consideration, PPMC using the odds ratio compares favorably with the use of Bayes factors (Raftery, 1996) and Pseudo-Bayes factors (Gelfand, 1996), which are more traditional Bayesian techniques for model comparison and selection. Sahu (2002) also found PPMC to be useful in facilitating model choice. Li et al. (in press) argued that PPMC can be potentially more useful because measures of item, person, and model fit used in PPMC can be more informative than (Pseudo-) Bayes factors, as PPMC can more precisely indicate where models fail and succeed. Though not pursued in the current work, PPMC for model comparison may be a natural extension.

Drawing from its appeal in traditional approaches, Karabatsos and Sheu (2004) employed a variant of Q_3 to conduct PPMC for local dependence in nonparametric item response models. In the context of PPMC for differential item functioning, Hoijtink (2001) developed a discrepancy measure to assess conditional independence at the model fit level.

In a series of simulation and applied studies, Sinharay and colleagues (Sinharay, 2005, in pressa, in pressb; Sinharay, Almond, & Yan, 2004; Sinharay & Johnson, 2003; Sinharay et al., in press) examined a variety of PPMC functions and techniques including (among others) direct data display, use of observed score distributions, odds ratios, Mantel-Haenszel statistics, and X^2 and G^2 measures at the item and model level to diagnose data-model misfit in IRT, LCA, BNs, and cognitive diagnosis models. The findings relevant to this study include the repeated success of the odds ratio and Mantel-

Haenszel statistics for sensing data-model misfit due to inadequacies in modeling dependence in the data (see especially Sinharay, 2005; Sinharay et al., in press).

In closing this section, we note that PPMC techniques have been applied to other psychometric models including SEM, LCA, and multilevel modeling, most often at the model level (Hoijtink, 1998; Hoijtink & Molenaar, 1997; Rubin & Stern, 1994; Scheines et al., 1999; Sinharay & Stern, 2003). No study has systematically examined the behavior of these measures as tools for PPMC under explicit hypotheses of the multidimensional structures of interest here as they apply to IRT or any psychometric model. The proposed work seeks to study this phenomenon in the case of certain IRT and BN models. McDonald and Mok (1995) argued that issues and questions of model fit cut across traditional bounds of IRT and SEM. To this we may add LCA, cognitive diagnosis, and multilevel models. Thus this work has the explicit intention of drawing conclusions with an eye towards application in psychometric models generally.

CHAPTER 3: CONDITIONAL COVARIANCE THEORY FOR MULTIDIMENSIONALITY

This chapter provides background for understanding the interplay between dimensionality and model fitting, motivating the design of the studies and several hypotheses regarding the influences of manipulated factors. More specifically, dimensionality assessment is couched in conditional covariance theory (Bolt, 2001; Stout et al., 1996; Zhang & Stout, 1999a). Factors hypothesized to influence dimensionality and dimensionality assessment are introduced in the context of this framework.

If weak local independence holds, conditioning on the latent variable(s) renders the conditional covariance between item-pairs to be zero. As a consequence, conditional covariances at or near zero are indicative of an adequately specified latent space. Nonzero conditional covariances are indicative of an underspecified latent space. In the balance of this chapter, we discuss factors that relate to the sign and magnitudes of the conditional covariance between pairings of items.

In working with multidimensional models, indeterminacies exist with regard to the estimation and representation of the latent dimensions. In the current work, we do not estimate multidimensional models (see Davey et al., 1996 for strategies to manage indeterminacies in estimation). In representing multidimensional models, we seek to employ representations that mimic the substantive concern for situations in which all items reflect a primary dimension but some items also reflect auxiliary dimensions. As discussed in this chapter, these representations allow us to leverage the results of the geometry of conditional covariance theory to explore features of multidimensional items.

GEOMETRY OF COMPENSATORY MULTIDIMENSIONAL TESTS

A number of graphical techniques are useful for representing multidimensional tests (Ackerman, 1996). We follow Stout et al. (1996) in using augmented versions of geometric representations developed by Reckase (1985; Reckase & McKinley, 1991; see also Ackerman, 1994; 1996) to represent the multidimensional structure of tests (Bolt, 2001; Stout et al., 1996; Zhang & Stout, 1999a). In passing we note the connections to geometric representations of linear factor analysis (Thurstone, 1947). Indeed, Zhang and Stout (1999a) provided a rigorous treatment of the geometry of conditional covariance theory for generalized compensatory multidimensional models, a class of models closely related to linear factor analysis. Key points from this line of research are reviewed here, motivating several hypotheses of interest in this work.

Let M=3 be the true dimension of a test. That is, for a set of J items, suppose there is 3-dimensional latent vector $\boldsymbol{\theta}_M = (\theta_1, \theta_2, \theta_3)$ that renders the J items locally independent and the removal of any of the M dimensions renders at least some items to be locally dependent. Further let each item follow a compensatory MIRT model.[3] Each item may be represented as a vector in 3-dimensional coordinate space where the vector is determined by the item parameters (Reckase, 1985, Zhang & Stout, 1999a). Item vectors for a M=3 dimensional test are given in Panel 1 of Figure 4 in the latent space defined by the latent variables, where the intersection of the axes may be thought of as the origin in the latent space. The position and orientation of an item vector is determined by its parameters (Reckase, 1985; Ackerman, 1996). For the current work, we ignore the effects of the location parameters and focus on the influence of the

[3] Zhang and Stout (1999a) provided theoretical results for *generalized* compensatory models, a class which includes the compensatory model as described by Equation (3). Thus the results reviewed here are applicable, but not limited, to the compensatory MIRT models pursued in this work.

discrimination parameters that capture which and how the latent variables influence the item.

The direction of each vector corresponds to the dimension(s) the item reflects, as determined by the discrimination parameters. The direction corresponds to the direction of best measurement for each item in the sense that this direction for an item is the composite of the latent dimensions along which the item provides maximal discrimination (Ackerman 1996; Reckase, 1985, Reckase & McKinley, 1991). For example, items X_1 and X_2 lie roughly evenly between θ_1 and θ_2, indicating that these items reflect θ_1 and θ_2 evenly. However, they do not vary along θ_3, indicating they do not reflect the third dimension.[4] Similarly, items X_3 and X_4 reflect θ_1 and θ_3 but not θ_2. The remaining items X_5 and X_6 are quite close to θ_1, and do not vary much along θ_2 or θ_3. To enhance visual distinguishability in this and later Figures, vectors for items reflecting θ_1 and θ_2 are plotted in red and vectors for items reflecting θ_1 and θ_3 are plotted in blue.

Though all items reflect θ_1, conditioning on θ_1 is not sufficient to render all pairs of items locally independent. The conditional covariances for certain pairs of items are nonzero. A key result from Zhang and Stout (1999a) is the connection between the conditional covariance for a pair of items conditional on some composite of the dimensions and the angle between the vectors resulting from projecting the item vectors

[4] Technically, both items X_1 and X_2 reflect the third dimension to a minimal extent. In the balance of this chapter, we refer to items in terms of the dimension(s) they *principally* reflect. Minimal variation along *other* dimensions is introduced for two reasons. First, having item vectors that lie exactly on top of one other makes it difficult to depict the relationships between the items in various situations, so some scatter is useful visually. Second, we do not wish to be restrictive, as the arguments of conditional covariance theory permit items to vary along all dimensions.

in *M*-dimensional space to the $(M-1)$-dimensional hyperplane orthogonal to the composite. For the items in Figure 4, if we condition on θ_1 (i.e., the composite is simply θ_1), the orthogonal hyperplane, denoted by $\perp\theta_1$, is the $\{\theta_2, \theta_3\}$ plane. Figure 4, Panel 2 shows the projections of the item vectors onto this plane. These projections in the 2-dimensional space of $\perp\theta_1$ are shown in Panel 3 of Figure 4. The use of the superscript '$\perp$' denotes that the items are projected into the plane orthogonal to the conditioned composite.

Figure 4: Item vectors in 3-dimensional latent space.

Panel 1

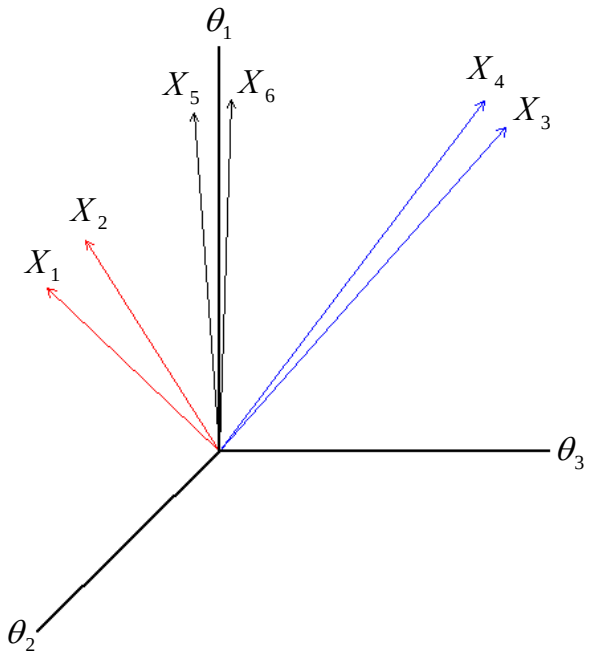
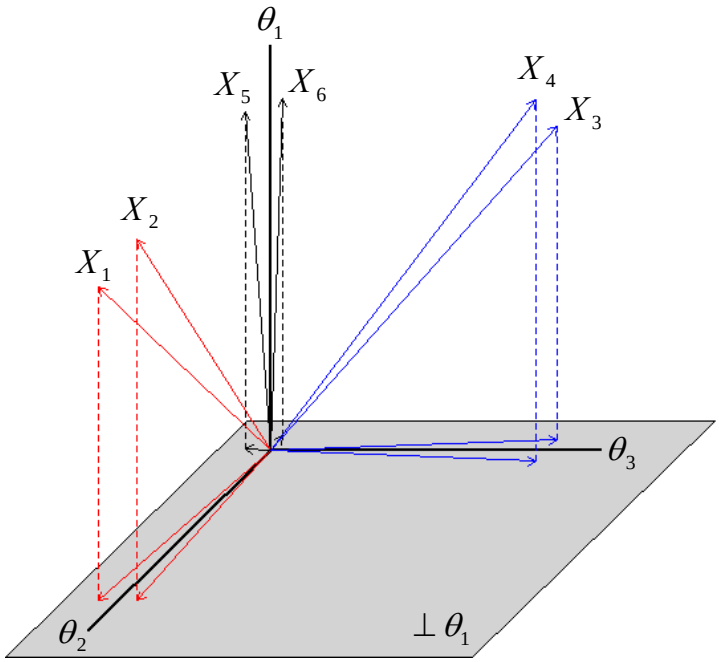
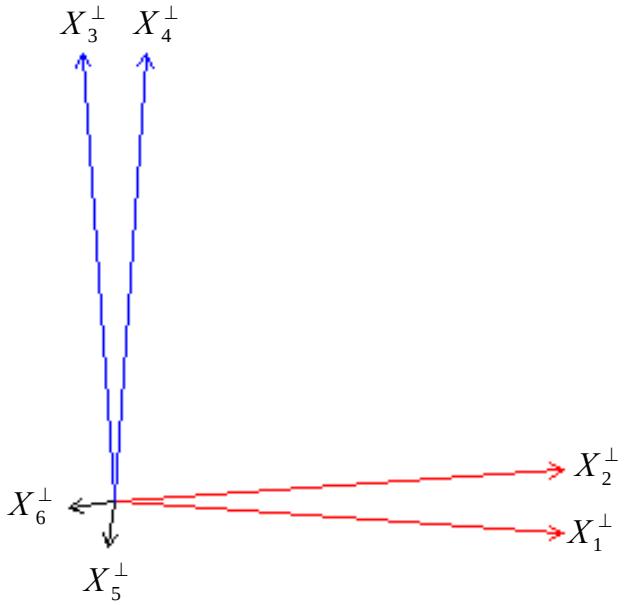


Figure 4 (continued): Item vectors in 3-dimensional latent space.

Panel 2



Panel 3



Zhang and Stout (1999a; Theorem 2, Equations (12)-(14)) proved that, provided that neither item is in the same direction of the composite, the conditional covariance between two items X_j and $X_{j'}$ is:

> 0 if the angle between $X_j^{\perp}$ and $X_{j'}^{\perp}$ is less than $\pi/2$;

= 0 if the angle between $X_j^{\perp}$ and $X_{j'}^{\perp}$ equals $\pi/2$;

< 0 if the angle between $X_j^{\perp}$ and $X_{j'}^{\perp}$ is greater than $\pi/2$.

Thus the size of the angle relative to $\pi/2$ determines the sign of the conditional covariance. Furthermore, the magnitude of the conditional covariance also depends on the angle. For *M*>2, holding the lengths of the item vectors and their position relative to the conditioned composite constant, as the angle between the projected vectors decreases (from π to 0), the conditional covariance increases (Zhang & Stout, 1999a). In other words, all else being equal, the smaller the angle, the larger the conditional covariance.

Returning to the items in Figure 4, this implies that the conditional covariance between items X_1 and X_2 is positive. Likewise, the conditional covariance between items X_3 and X_4 is positive. The conditional covariance between item X_1 (or item X_2) and item X_3 (or item X_4) is close to zero, as the angle between the projected vectors is close to $\pi/2$.

Zhang and Stout (1999a) further conjectured that, holding all else constant, the further away the items are from the composite, the larger the conditional covariance. When one or both of the items are close to the composite, the smaller the conditional covariance. Hence, though the angle between $X_1^{\perp}$ and $X_6^{\perp}$ is greater than $\pi/2$ (*prima facie* evidence of a negative conditional covariance), the conditional covariance between

items X_1 and X_6 is not far from zero, as $X_6^{\perp}$ is very close to the composite (Figure 4, Panel 3). Similarly, all the conditional covariances between any of items $X_1, \ldots, X_4$ and either of X_5 and X_6 ought to be close to zero.

FACTORS AFFECTING LOCAL DEPENDENCE

Structures like those in Figure 4, in which all items reflect the primary dimension (θ_1) and some items reflect an auxiliary dimension (θ_2 or θ_3), are the focus of this study. Examples of situations in which all test items reflect a single dimension of inferential interest but certain items also reflect auxiliary dimensions include tests that exhibit testlet effects, rater effects, or differential item functioning (e.g., Bolt & Stout, 1996; Bradlow et al., 1999, Shealy & Stout, 1993; Yen, 1993)

As just discussed, conditioning on θ_1 (or any other composite) in such situations is not sufficient to render all items conditionally independent. In this section, factors affecting the amount of local dependence between items are explicated conceptually and geometrically.

Strength of Dependence on Auxiliary Dimensions

To the degree that items more strongly reflect an auxiliary dimension, the less the primary dimension is able to fully account for the covariance between the items. Items that reflect both θ_1 and θ_2 contain two sources of association. Conditioning on just one, θ_1, accounts for the association due to θ_1 but not that due to θ_2. The stronger the dependence of the items on θ_2, the stronger the items are associated after conditioning on θ_1. Likewise for items that reflect both θ_1 and θ_3.

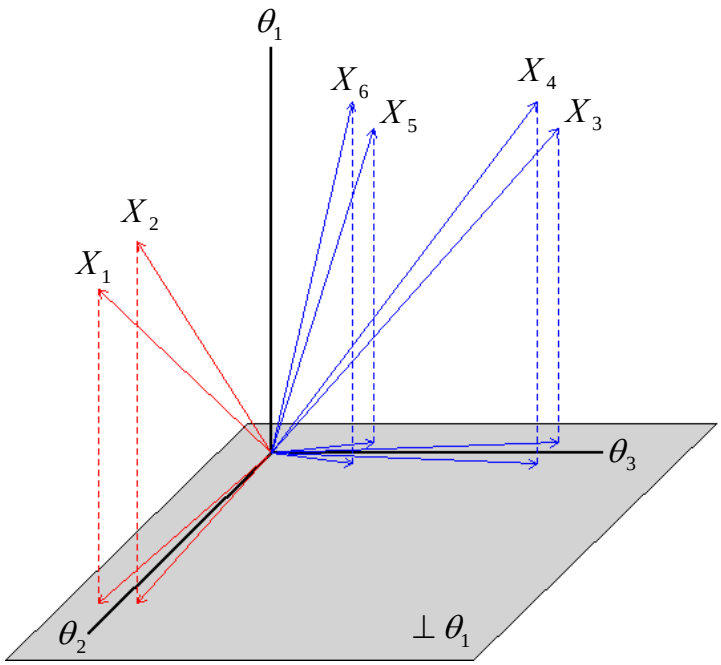
The strength of dependence of a compensatory MIRT item on a dimension is captured by the discrimination parameter for the item on the dimension (a_{jm} in Equation (3)). Akin to loadings in factor analysis, the a_{jm} reflect the pattern and strength of dependencies of the observable variables on the latent variables. Theorem 2, Equation (12) of Zhang and Stout (1999a) suggests that an increase in the discrimination parameter for any dimension that is in a direction other than the dimension conditioned upon leads to an increase in the magnitude of the conditional covariance. In the current analysis, θ_1 is being conditioned upon. Increases in the absolute value of the discrimination parameters for θ_2 and θ_3 lead to increases in magnitude of the conditional covariance.

To illustrate, Figure 5, Panel 1 depicts a set of item vectors. Items X_3 and X_4 reflect θ_1 and θ_3 roughly evenly. Items X_5 and X_6 also reflect θ_1 and θ_3 . They reflect θ_1 to the same degree as do items X_3 and X_4, respectively. However, the strengths of their dependence on θ_3 are one third those of X_3 and X_4 on θ_3, respectively:

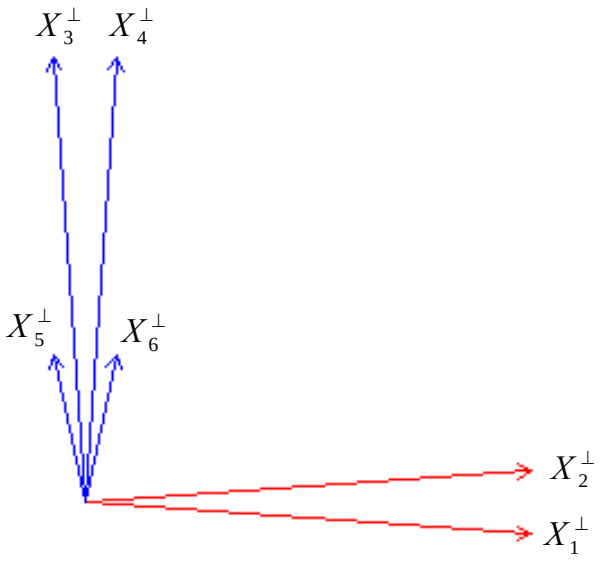
$$a_{51} = a_{31}$$
$$a_{61} = a_{41}$$
$$a_{53} = \frac{a_{33}}{3}$$
$$a_{63} = \frac{a_{43}}{3}.$$

Figure 5: Item vectors in 3-dimensional latent space illustrating the strength of dependence on auxiliary dimensions.

Panel 1



Panel 2



Panel 2 of Figure 5 plots the projected item vectors in the plane orthogonal to θ_1. The angle between $X_5^{\perp}$ and $X_6^{\perp}$ is larger than the angle between $X_3^{\perp}$ and $X_4^{\perp}$. Accordingly, the conditional covariance between X_5 and X_6 is less than the conditional covariance between X_3 and X_4 (Zhang & Stout, 1999a). Note that the angle between X_5 and X_6 in latent variable space (in Panel 1) is identical to that of X_3 and X_4. Thinking in terms of pairs of items, the shift from the pairing of X_3 and X_4 to the pairing of X_5 and X_6 is a reduction in the dependence on θ_3, but not on θ_1. Geometrically, in moving from pairing of X_3 and X_4 to the pairing of X_5 and X_6, we have rotated the item vectors closer to θ_1, the dimension being conditioned on. Figure 5, Panel 2 shows that this manifests itself in a larger angle between the projected item vectors and hence a smaller conditional covariance. This result supports the conjecture of Zhang and Stout (1999a) that, holding the angle between the item vectors constant, the conditional covariance increases as the items are rotated away from the conditioning dimension.

Correlations Among Dimensions

A MIRT model specifies that the associations amongst a set of variables are attributable to their dependencies on common latent dimensions. In addition to dependencies between items, the latent dimensions themselves may be correlated. The influence of the correlations among the latent dimensions on the conditional covariances between the items is explored in this section.

Geometrically, the angle between the dimensions captures the correlations among them. Let $\rho_{mm'}$ denote the correlation between dimensions θ_m and $\theta_{m'}$. The geometry of multidimensional structures implies that the angle between θ_m and $\theta_{m'}$ is (Zhang & Stout, 1999a)

$$< \frac{\pi}{2} \text{ if } \rho_{mm'} > 0$$

$$= \frac{\pi}{2} \text{ if } \rho_{mm'} = 0$$

$$> \frac{\pi}{2} \text{ if } \rho_{mm'} < 0.$$

For ease of exposition, we restrict attention to situations in which the correlations between the dimensions are equal. Figure 6 depicts dimensions θ_2 and θ_3 as uncorrelated with θ_1. Dimensions θ_2^*, and θ_3^* are correlated with each other and with θ_1; each bivariate correlation is .6. Accordingly, the angles between these dimensions are less than $\pi/2$. For visual perspective, the dashed lines show the projections θ_2^* and θ_3^* into the $\{\theta_2, \theta_3\}$ plane (note that each dimension is plotted to be of equal length in 3-dimensional space). We may think of θ_2^* and θ_3^* as transformations of θ_2 and θ_3, respectively. As the correlations between the dimensions increases from zero to .6, θ_2 and θ_3 rotate toward each other and toward θ_1, ultimately becoming θ_2^* and θ_3^*, respectively.[5]

[5] Note that in the representation adopted here, θ_1 remains unmoved. This is an arbitrary choice in solving a rotational indeterminacy.

Figure 6: Uncorrelated and correlated dimensions.

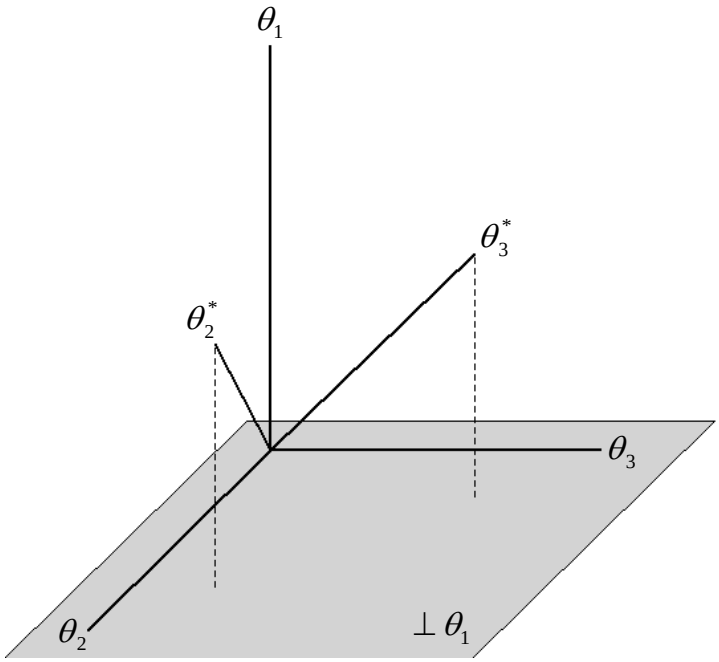
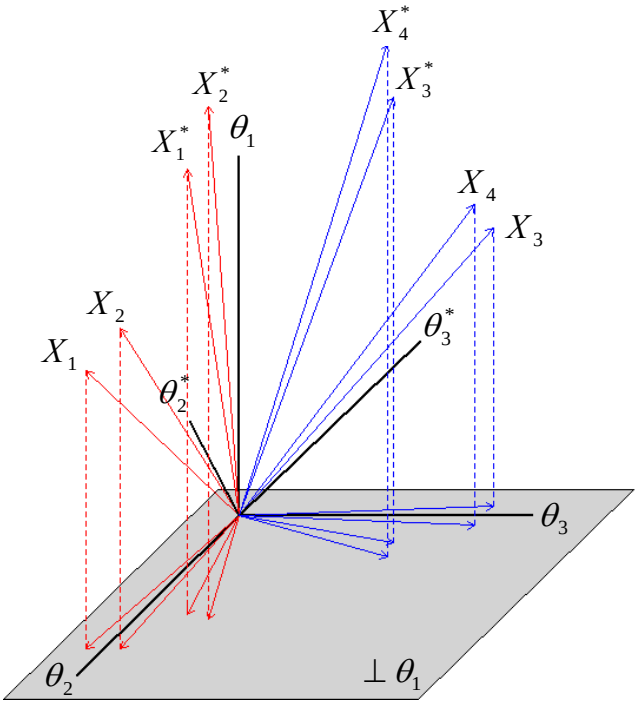


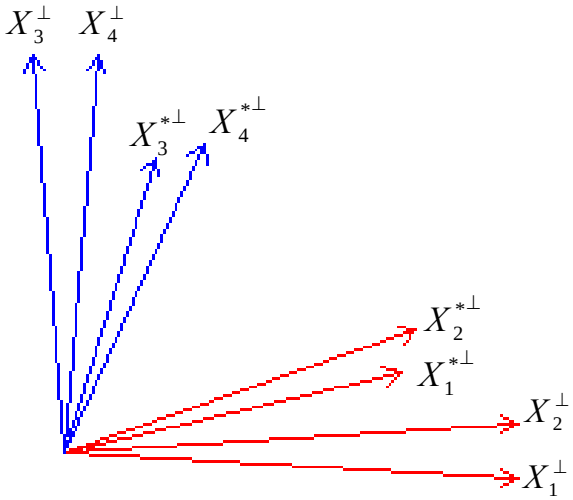
Figure 7, Panel 1 adds plots of item vectors in the space. Items X_1 and X_2 reflect dimensions θ_1 and θ_2, and items X_3 and X_4 reflect dimensions θ_1 and θ_3. As just discussed, when the correlations between the dimensions increase to .6, θ_2 and θ_3 become θ_2^* and θ_3^*. Holding all else constant, items X_1^*, X_2^*, X_3^*, and X_4^* are the resulting transformations of items X_1, X_2, X_3, and X_4, respectively. As can be seen by their projections onto $\perp \theta_1$, the result of the positive correlations are item vectors that are more closely aligned to θ_1. Figure 7, Panel 2 plots the projected item vectors in $\perp \theta_1$ and makes this last point more explicit. The projected vectors for the items based on the correlated dimensions do not extend as far as their counterparts from the uncorrelated dimensions.

Figure 7: Item vectors for uncorrelated and correlated dimensions.

Panel 1



Panel 2



The angle between $X_1^{*\perp}$ and $X_2^{*\perp}$ is slightly smaller than the angle between $X_1^{\perp}$ and $X_2^{\perp}$, which suggests that the conditional covariance between $X_1^{*\perp}$ and $X_2^{*\perp}$ is larger than that between $X_1^{\perp}$ and $X_2^{\perp}$ (Zhang & Stout, 1999a). However, it is advanced here that whatever increase in conditional covariance due to a decrease in angle is more than offset by the rotation of the items closer to the conditioning dimension, which leads to the lengths of the vectors $X_1^{*\perp}$ and $X_2^{*\perp}$ being shorter than their counterparts $X_1^{\perp}$ and $X_2^{\perp}$. The net result is that the conditional covariance between X_1^* and X_2^* is less than the conditional covariance between X_1 and X_2. Likewise for X_3^* and X_4^* relative to X_3 and X_4.

Conceptually, as the correlations between the dimensions increase, the dimensions become more collinear. Holding item parameters constant, the result is that the trajectories of the items become more collinear with themselves and with the conditioned upon dimension, θ_1. This is clearly seen in Figure 7, Panel 1, in which the trajectories of items $X_1^*,\ldots,X_4^*$ are closer to θ_1 (and each other) than are the trajectories of items $X_1,\ldots,X_4$. In the limit, as the correlations among the dimensions go to unity, the trajectories of all items go to the trajectory of θ_1.

Mathematically, when (at least) one item has a trajectory in the same direction as the conditioned upon dimension, the conditional covariance is zero (Zhang & Stout, 1999a). Conceptually, when all the dimensions become θ_1, then conditioning upon θ_1 is sufficient to explain all the associations among the items. This is clearly seen in the compensatory MIRT model studied here (Equation (3)). When the correlations between

the dimensions are all unity, the model reduces to a unidimensional (i.e., 2-PL) model in which the discrimination parameter for each item is the sum of the item's multidimensional discrimination parameters.

Proportion of Items Exhibiting Multidimensionality

The preceding analyses have examined the behavior of the conditional covariance and its geometric representation upon conditioning on θ_1. In practice, we are rarely able to condition on a latent dimension *a priori*. Rather, when fitting a model, we *estimate* the dimension concurrently with estimation of item parameters. A notable exception exists when items are calibrated with examinees that have known (or treated as known) values along the dimension, as is the case when pilot items are embedded in an operational assessment. Values for examinees on the latent variable(s) are treated as known in estimating the item parameters. Conditioning on the latent variable(s) to analyze the conditional association of the variables may then be possible. The more general situation, and the one assumed for the balance of this section, is one in which examinee variables and item parameters are estimated simultaneously.

In calibrating a unidimensional model a single latent dimension is estimated. This dimension is the dimension of "best measurement" in the sense that it is the dimension along which the items maximally discriminate (Reckase, 1985; Reckase & McKinley, 1991).

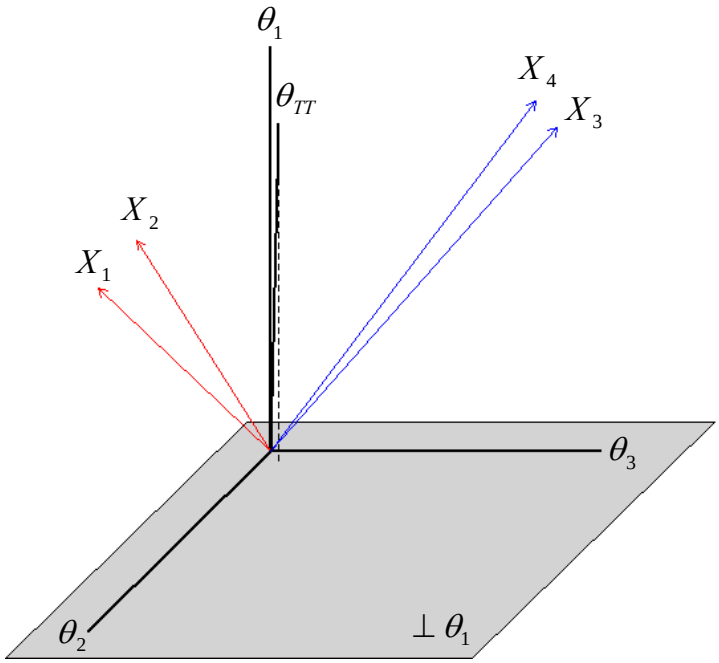
Reckase (1985; Reckase & McKinley, 1991) provided foundational work on understanding the dimension of best measurement in terms of a direction in multidimensional space. Advances in geometric and other graphical analyses in MIRT (Ackerman, 1994; 1996) led to theoretical developments in understanding the dimension

(direction) of best measurement for a set of items and the subsequent influence on the conditional associations among multidimensional variables conditional on this dimension (Bolt, 2001; Stout et al., 1996; Zhang & Stout, 1999a). Estimating this best dimension and then conditioning upon the estimate of the dimension are key components in a number of tools for assessing dimensionality and related psychometric phenomena (Bolt & Stout, 1996; Shealy & Stout, 1993; Stout et al., Zhang & Stout, 1999a).

Returning to the current research, the relevant result is that when the data exhibit multidimensionality, it is *not* the case that the single best dimension will be the primary dimension of inferential interest (θ_1). Rather, the estimated dimension will be some composite of the true, underlying latent dimensions (Reckase, 1985; Reckase & McKinley, Zhang & Stout, 1999a).

The best dimension of measurement may be represented geometrically as an appropriate combination of the items comprising the test (Stout et al., 1996). When almost all the items measure the primary dimension only, and just a few items reflect multiple dimensions, the direction of best measurement will be quite close to the primary dimension. This is depicted in Figure 8. There are 28 items (not shown) that principally reflect θ_1 only. As in previous figures, X_1 and X_2 reflect both θ_1 and θ_2 roughly equally; X_3 and X_4 reflect θ_1 and θ_3 roughly equally. In this situation, a low proportion (four of 32) items reflect dimensions other than θ_1. The dimension best measured by the total test, θ_{TT}, is quite close to θ_1. This is evident by projecting θ_{TT} into the plane orthogonal to θ_1 in which it is observed that this projection intersects $\perp\theta_1$ near the origin.

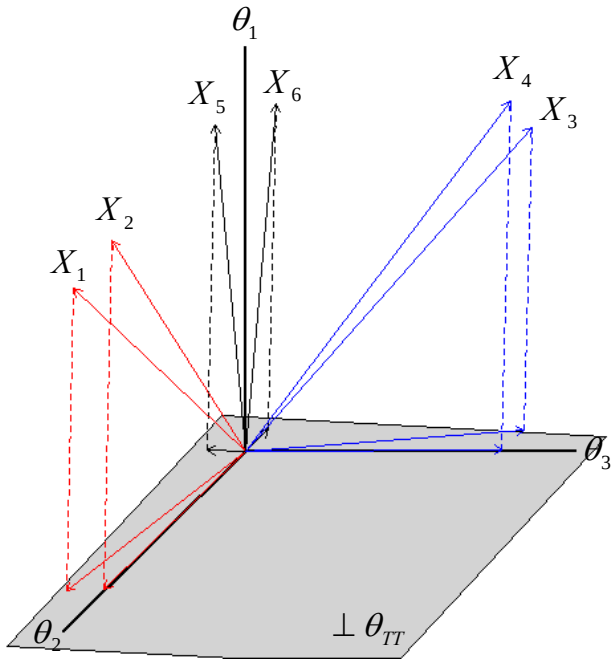
Figure 8: Dimension of best measurement with a low proportion of multidimensional items.



Model calibration results in the estimated dimension being θ_{TT}, which implies that θ_{TT} is the dimension that should be conditioned upon. Figure 9, Panel 1 shows the projections of the vectors for items $X_1,\ldots,X_4$ into the plane orthogonal to θ_{TT}, denoted as $\perp\theta_{TT}$. Note that $\perp\theta_{TT}$ is quite close, but not identical, to $\perp\theta_1$ in Figure 8 (it is slightly tilted down from $\perp\theta_1$). Thus the projected item vectors are quite similar to what they would have been if θ_1 was conditioned upon. In addition, two other items, X_5 and X_6 are plotted along with their projections. These items are representative of the 28 items that reflect θ_1 only. Figure 9, Panel 2 plots the projected item vectors in $\perp\theta_{TT}$.

Figure 9: Item vectors projected into $\perp \theta_{TT}$ with a low proportion of multidimensional items.

Panel 1



Panel 2

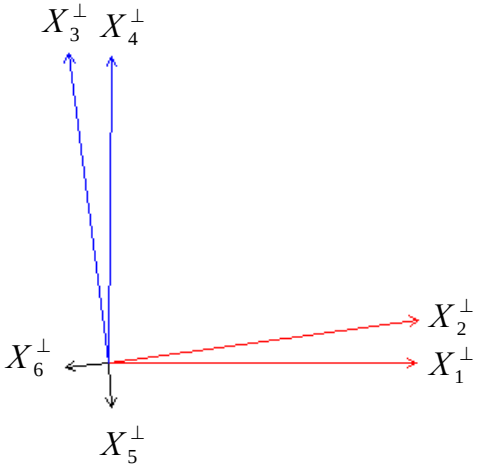


Figure 10 depicts the change in θ_{TT} when the proportion of items reflecting multiple dimensions increases. In this case, there are only four items (not pictured) that principally reflect θ_1. Fourteen items principally reflect θ_1 and θ_2; X_1 and X_2 are two such items. Likewise X_3 and X_4 represent 14 items that principally reflect θ_1 and θ_3. In the previous example, the vast majority of the 32 items reflected θ_1 only. In this example, a few items reflect θ_1 and the vast majority reflect multiple dimensions (half of these reflect θ_2, the other half reflect θ_3). The effect on θ_{TT} is considerable. As its projection into $\perp\theta_1$ reveals, it is now much further from θ_1 than when the proportion of multidimensional items is low (Figure 8).

Figure 10: Dimension of best measurement with a high proportion of multidimensional items.

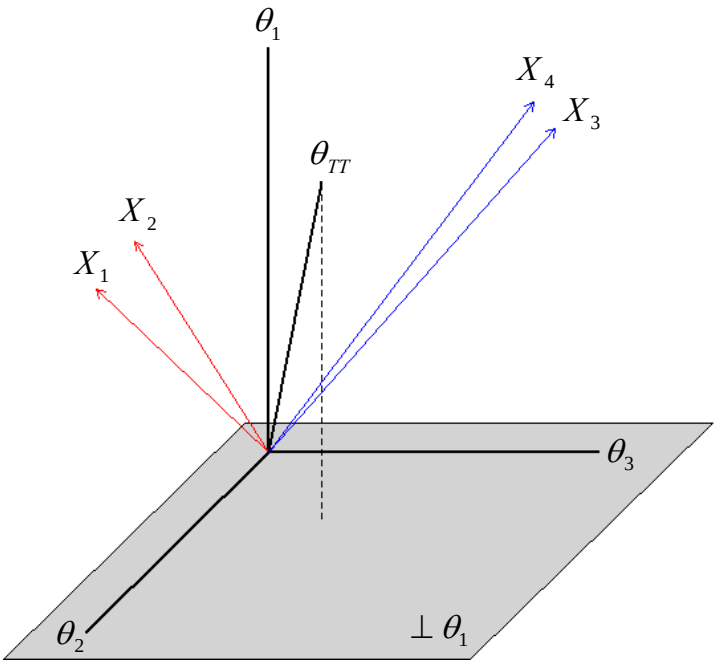
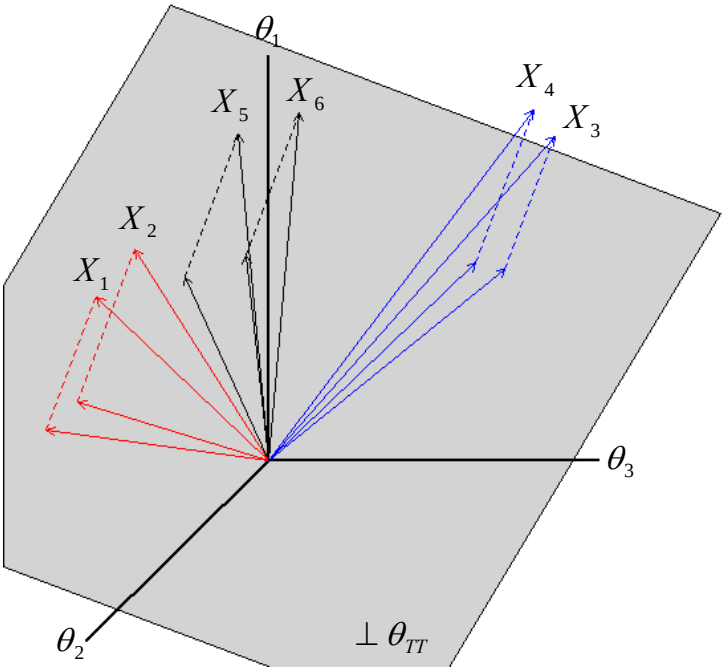
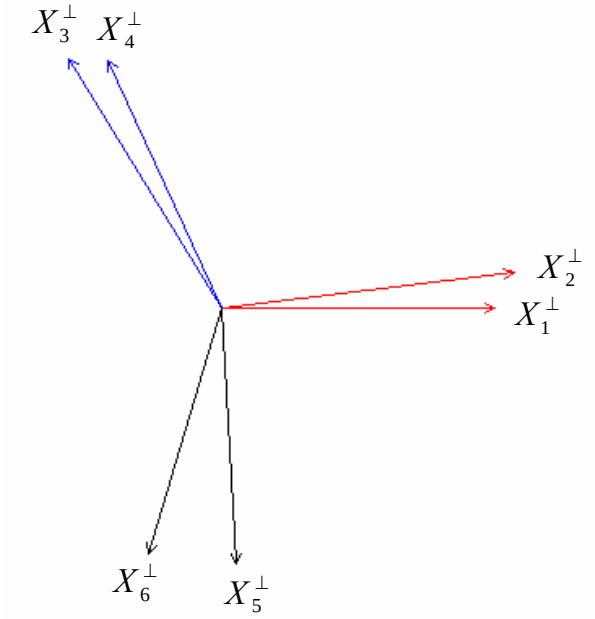


Figure 11: Item vectors projected into $\perp \theta_{TT}$ with a high proportion of multidimensional items.

Panel 1



Panel 2



As a consequence of the departure of θ_{TT} from θ_1, $\perp\theta_{TT}$ becomes much more tilted relative to $\perp\theta_1$. Panel 1 of Figure 11 shows this plane and the projections of six items. Panel 2 depicts these projections in $\perp\theta_{TT}$. The six items are the same as those in the previous analysis; all that has changed is the dimensionality of the remaining items. Comparing Figures 8 and 10, it is observed that as the proportion of multidimensional items increases, θ_{TT} moves away from θ_1. The plane $\perp\theta_{TT}$ also shifts (Panel 1 of Figures 9 and 11), as do the projections of the item vectors (Panel 2 of Figures 9 and 11).

Beginning with the items that the same reflect multiple dimensions, the angle between $X_1^{\perp}$ and $X_2^{\perp}$ (likewise, the angle between $X_3^{\perp}$ and $X_4^{\perp}$) has not changed meaningfully. However, the length of each of these vectors is shorter, indicating a reduction in the conditional covariance.

The angles that have changed meaningfully are those between items that reflect *different* multiple dimensions. For example, the angle between $X_1^{\perp}$ and $X_3^{\perp}$ is slightly greater than $\pi/2$ when the proportion of items reflecting multiple dimension is low (Figure 9) but is noticeably larger when the proportion of items reflecting multiple dimensions is high (Figure 11). Angles larger than $\pi/2$ imply *negative* conditional covariances (Zhang & Stout, 1999a). Habing and Roussos (2003) proved that, if at least two items exhibit positive local dependence with respect to $\perp\theta_{TT}$, there must be at least one pairing of items that exhibits negative local dependence. That is, in addition to underpredicting the associations between items that reflect the same auxiliary dimension, the model overpredicts the associations between items that reflect different auxiliary dimensions (Chen & Thissen, 1997; Sinharay, et al., in press). This manifests itself in the

conditional covariances for items reflecting different auxiliary dimensions being negative, represented by the obtuse angles between such item-pairs.

Another effect of increasing the proportion of multidimensional items is the increase in the length of the vectors for $X_5^{\perp}$ and $X_6^{\perp}$ and the decrease in the angle between them, indicative of an increase in their conditional dependence. This effect can be explained conceptually as follows. When there are only a few multidimensional items, θ_{TT} is close to θ_1; items that reflect θ_1 only vary around θ_{TT} minimally. As the proportion of multidimensional items increases, θ_{TT} drifts farther away from θ_1. Hence the items that only reflect θ_1 are more distinct from θ_{TT}. They are more similar to each other than can be accounted for by θ_{TT}.

The movement of θ_{TT} away from θ_1 also explains the other effects. As the proportion of multidimensional items increases, θ_{TT} moves toward the middle of the latent space in such a way that it is closer to both θ_2 and θ_3, hence reducing the conditional covariances for pairs of items that both reflect θ_2 or that both reflect θ_3. However, θ_{TT} lies *in between* θ_2 and θ_3. Accordingly, the item vectors are on opposite sides of θ_{TT} (Figures 9 and 11) and are hence *negatively* associated, conditional on θ_{TT}.

Strength of Dependence on Auxiliary Dimensions and Correlations Among the Latent Dimensions Revisited

The prior discussions on the influences of (a) the strength of dependence of items on the auxiliary dimensions and (b) the correlations among the latent variables were couched in a situation in which θ_1 was conditioned on. In practice, θ_1 is not known to

the analyst, and some estimate of θ_{TT} is employed. Nevertheless, arguments similar to those made regarding the influences of these factors based on conditioning on θ_1 could be made based on conditioning on θ_{TT}. Such arguments will not be presented in full. The balance of this section conceptually describes the influence of estimating and conditioning on θ_{TT} rather than θ_1 on the aforementioned factors.

Strength of Dependence on Auxiliary Dimensions

Increasing the strength of dependence of items on the auxiliary dimensions increases the variability of the *J*-variate distribution in the latent space. We hypothesize that the effect on θ_{TT} is that θ_{TT} will drift farther from θ_1. As the strength of dependence of an item on an auxiliary dimension increases, it "drags" θ_{TT} further into the multidimensional space, away from θ_1. As seen in the analysis of the proportion of multidimensional items this has the effect of reducing the conditional covariance between items that reflect the same multiple dimensions.

Thus, increasing the strength of dependence results in two *opposite* effects. The first is that as the strength of dependence on multiple dimensions increases, a single latent dimension becomes less and less capable of accounting for all the sources of association between items, which in turn leads to *increases* in the conditional associations. The second effect is on θ_{TT}; increasing the strength of dependence brings θ_{TT} further away from θ_1 and closer to the auxiliary dimensions. θ_{TT} is therefore better able to explain the associations for items that reflect the same multiple dimensions, which leads to *decreases* in the conditional associations for such pairs of items. We hypothesize that the first effect is stronger than the second and that in the main, increasing the strength of

dependence on the auxiliary dimensions leads to increases in the conditional associations for items that reflect the same multiple dimensions.

Correlations Among the Latent Dimensions

As argued above, increases in the correlations between the latent dimensions serve to effectively reduce the dimensionality of the latent space. As the correlations increase toward unity, the latent space goes to a single latent dimension. Accordingly, θ_{TT} ought to get *closer* to θ_1 as the correlation increases. Though the need to estimate θ_{TT} suppresses the effect of the strength of dependence, we hypothesize that it *enhances* the effect of increasing the correlation. In other words, larger (positive) correlations among the latent dimensions bring the items together and reduces the conditional covariances (Figure 7). Furthermore, θ_{TT} will be closer to the multidimensional items than θ_1, which also reduces the conditional covariances.

Implications for Study Design

We do not pursue the differences between conditioning on θ_1 as opposed to θ_{TT} in this study. In an effort to more closely mimic operational practice, we proceed with model estimation and model checking by estimating the latent variable along with the item parameters in calibrating the model.

CONJUNCTIVE MODELS

The preceding characterization of multidimensional models is based on foundational work on compensatory and generalized compensatory models (Reckase, 1985; Reckase & McKinley, 1991; Stout et al., 1996; Zhang & Stout, 1999a). No such

foundation has been established for conjunctive models. However, the factors discussed in this chapter are more general than their compensatory manifestations. Statements that the (a) strength of dependence on auxiliary dimensions, (b) correlations among the latent dimensions, and (c) proportion of multidimensional items all influence dimensionality assessment may be made independently of any particular form of the IRF. To this end, conceptual arguments (i.e., those not tied to any IRF) were advanced for each of the factors above.

The implication of this generality is that the same hypotheses may be advanced for a class of multidimensional models broader than those for which we may leverage the machinery of geometric representations. Accordingly, the stated hypotheses regarding the effects of the strength of dependence of items on auxiliary dimensions, the correlations among the dimensions, and the proportion of multidimensional items are hypothesized to hold for conjunctive models.

Operationally, the concepts of the correlations between the latent dimensions and the proportion of multidimensional items remain unchanged; they do not depend on the form of the IRF. What is needed is a mechanism for operationalizing notions of strength of dependence on auxiliary dimensions in conjunctive MIRT models.

The mechanism proposed is to employ the location parameter along a dimension 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.

SUMMARY

Conditional covariance theory provides a theoretical framework for understanding the interplay between multidimensional data and unidimensional models fit to such data. The theory also affords geometric representations useful for conveying key features relevant to conditional covariances. The framework has been leveraged to provide grounding for tools that assess unidimensionality (Nandakumar & Stout, 1993; Stout 1987), estimate the number of distinct, dimensionally homogeneous clusters of items (Zhang & Stout, 1999b), and model differential item functioning (Bolt & Stout, 1996; Shealy & Stout, 1993). Considering multidimensionality from a conditional covariance perspective in part motivates the design of these studies and the hypothesized effects.

CHAPTER 4: POSTERIOR PREDICITVE MODEL CHECKING FOR MULTIDIMENSIONALITY IN IRT

The purpose of this study is to investigate the use of PPMC for detecting misfit in IRT models due to the failure to adequately model the underlying dimensions. As reviewed above, PPMC has received an increasing amount of attention in the psychometric community, and even though many psychometric phenomena can be framed in terms of multidimensionality, research applying PPMC in such situations has been limited.

RESEARCH DESIGN

A Monte Carlo study is conducted in which a number of factors hypothesized to be relevant to detection of multidimensionality are varied. Multiple data sets are generated under null and non-null conditions to facilitate an examination of the application of PPMC to situations in which a unidimensional model is hypothesized, but the data exhibit multidimensionality.

All data sets consisted of J=32 items, consistent with typical test lengths in both research and operational assessments. The number of items was not varied because the proportion of items reflecting multiple dimensions was of interest. By keeping the total number of items fixed, the proportion of multidimensional items may be easily manipulated.

With the exception of null conditions, all data sets were generated from a model with three latent dimensions: θ_1, θ_2, and θ_3. All items were dependent on θ_1 and as discussed below, certain items reflect θ_2 in addition to θ_1 and other items reflect θ_3 in

addition to θ_1. For null conditions, data were generated from a model with a single latent dimension.

The distribution of the latent dimension is multivariate normal where all three dimensions have mean zero and variance one. As discussed below, correlations between the dimensions were manipulated.

Manipulated Factors

Five factors related to multidimensionality were manipulated in this study. Each factor and the associated levels are described in turn.

Data-Generating Model

Data were generated from both compensatory MIRT (Equation (3)) and conjunctive MIRT models (Equation (4)). Table 1 provides a layout of several of the conditions of the study. The first column is the item number and the second column gives the value of the location parameter in the data-generating MIRT model. The values in this column will be the values for d_j in the compensatory MIRT conditions and $-b_{j1}$ in the conjunctive MIRT conditions. Most items are concentrated near the center of the ability distribution, with a few towards the extremes, a pattern which is consistent with operational assessments.

For the compensatory MIRT conditions, the discrimination along the primary dimension, a_{j1}, is 1 for all items. This was done because a target of this research is the influence of the strength of dependence of the items on the different dimensions relative to their dependence on the primary dimension. By holding the discrimination along the

primary dimension constant, the strength of dependence could be manipulated by varying the discrimination parameters along the auxiliary dimensions, as described below.

Proportion of Multidimensional Items

The proportion of items reflecting multiple dimensions varied from low (four items) to medium (16 items) to high (28 items). Table 1 indicates which items reflect the auxiliary dimensions in these conditions. The mark 'X' indicates that the item reflects the second or third dimension. For simplicity, the only cases considered will be those in which an equal number of items depend on the second and third dimensions. Note that the items are listed in increasing difficulty and the pattern of multidimensionality is symmetric with respect to the center (i.e., symmetric around zero). This is done to insure that there is no correlation between marginal item difficulty and which dimension(s) the item reflects, so as to not confound difficulty and dimensionality (Ackerman, 1996). Had all the multidimensional items been concentrated at one end of the continuum, any patterns of misfit might have been attributable to item difficulty, rather than dimensionality. Similarly, balance between the auxiliary dimensions was maintained by allotting the same number of items to reflect θ_2 as reflect θ_3.

Table 1: Patterns of multidimensionality

		Number of items reflecting multiple dimensions					
		4		16		28	
Item	Location[a]	θ_2	θ_3	θ_2	θ_3	θ_2	θ_3
1	2.5						
2	2				X		X
3	1.75			X		X	
4	1.5						X
5	1.34					X	
6	1.17		X		X		X
7	1			X		X	
8	0.875						X
9	0.75					X	
10	0.625				X		X
11	0.5	X		X		X	
12	0.4						X
13	0.3					X	
14	0.2				X		X
15	0.1			X		X	
16	0						
17	0						
18	-0.1			X		X	
19	-0.2				X		X

20	-0.3						X	
21	-0.4							X
22	-0.5	X			X		X	
23	-0.625					X		X
24	-0.75						X	
25	-0.875							X
26	-1				X		X	
27	-1.17			X		X		X
28	-1.34						X	
29	-1.5							X
30	-1.75				X		X	
31	-2					X		X
32	-2.5							

a. For compensatory MIRT, the location is d_j. For conjunctive MIRT, the location is $-b_{j1}$.

Strength of Dependence

For compensatory MIRT data, the degree to which item performance depends on the second (or third) dimension is captured by a_{j2} (or a_{j3}). In each analysis, the values for a_{j2} (for items that reflect the second dimension) will be equal to that of a_{j3} (for items that reflect the third dimension). The values of a_{j2} and a_{j3} are varied from weak (.25) to moderate (.5) to strong (.75) to equal (1). These are interpreted relative to a_{j1}, which is 1 for all items. Thus, with the variances of the latent variables equal to each

other the values correspond to an influence of the second or third dimension that ranges from one-quarter to equal that of the influence of the first dimension.

For conjunctive MIRT data, the degree to which item performance depends on the second (or third) dimension is captured by b_{j2} (or b_{j3}). The values for b_{j2} and b_{j3} (constant over items in any one condition) are varied from -1.0 to -0.5 to 0.5 to 1.0. Because the auxiliary dimensions are marginally distributed as standard normal variables, a b_{j2} (or b_{j3}) value of -1.0 models the case where (approximately) 84% of the examinees have more than a 50% chance of "passing" on the second (or third) dimension. The values of -0.5, 0.5, and 1.0 correspond to approximately 70%, 30%, and 16% (respectively) of the examinees having more than a 50% chance of "passing" on the second (or third) and dimension.

Correlations Between the Latent Dimensions

For simplicity the bivariate correlations between the latent dimensions were always equal. The chosen values for the correlations between span the range from no association (0) to weak (.3) to strong (.7) to extreme (.9). The limiting case of 1.0 is not included, as when the correlation is 1.0, the three-dimensional latent space reduces to one dimension.

Sample Size

The three conditions for sample size are 250, 750, and 2500 examinees.

Modeling

There are 288 combinations of the manipulated factors. In addition, null conditions of unidimensionality were investigated in which the data are generated from a

compensatory MIRT model with the discriminations along the second and third dimensions set to zero. Three additional null cells corresponding to the three sample sizes are included, yielding a total of 291 cells.

Model Fitting

For each cell, 50 replications of the following procedures were conducted. Data were generated according to the model specified by the cell, and the 2-PL model was estimated with the following prior distributions:

$$\theta_i \sim N(0,1) \quad i = 1,\ldots,N;$$

$$b_i \sim N(0,1) \quad j = 1,\ldots,J;$$

$$\ln(a_i) \sim N(0,1) \quad j = 1,\ldots,J.$$

The choice of these priors represents knowledge and restrictions brought to bear on the analysis. The use of normal priors for examinee latent variables and item difficulty parameters allow these variables to span the real line. The use of lognormal priors for item discrimination restrict these variables to be positive, imposing the restriction that all items be positively related to θ. The choices of the distributional forms and the values of parameters for the priors reflect common choices for logistic IRT models (e.g., Mislevy, 1986a; Rupp et al., 2004). The use of common, independent prior distributions for examinee variables (likewise item parameters) reflects assumptions of exchangeability (de Finetti, 1964; Lindley & Novick, 1981; Lindley & Smith, 1972) regarding examinees (likewise items). Though the variables are specified as independent in the prior distributions, they are not restricted to be independent in the posterior, should the data suggest dependencies (Gelman, 1996).

Estimation

The MCMC technique utilized is a Metropolis-within-Gibbs sampler (Patz & Junker, 1999a) in which the parameters are updated univariately based on their full conditional distributions (the Gibbs component). Since the full conditional distributions are not tractable (Maris & Bechger, in preparation) a Metropolis step is taken within each Gibbs step. A random draw is taken from a (univariate) normal proposal distribution and this candidate point is accepted as the next value for the parameter with probability defined by the ratio of the heights of the candidate and current point in the posterior distribution (Chib & Greenberg, 1995). If the candidate is not accepted, the current value is retained as the next value in the chain. A more complete description of the Metropolis-within-Gibbs MCMC algorithm is given in Appendix A.

Pilot studies were conducted to determine near optimal values for the variance of the proposal distributions used in the sampler. Judicious choices for the proposal distributions can result in faster convergence and improved mixing. On the basis of this pilot work (summarized in Appendix A), the number of iterations discarded as burn-in was 500 for conditions in which the sample size was 250 or 750. For conditions with a sample size of 2500, 600 iterations were discarded as burn-in.

For each analysis, five chains were run in parallel from overdispersed starting points (Brooks & Gelman, 1998; Gelman & Rubin, 1992). Each chain was run for 200 iterations after burn-in. These 200 iterations were thinned by 2 to mitigate the effects of serial dependencies within a chain. The resulting 100 iterations from each chain were pooled, totaling 500 iterations to be used in PPMC. The use of multiple chains serves to reduce the dependencies among the iterations, as though there may be dependencies

between iterations *within* a chain, there are no dependencies *between* chains (Gelman, 1996).

The Metropolis-within-Gibbs sampling scheme (Appendix A) was programmed in C++.[6] For the large sample size condition, to complete one cell it took about 14 hours on a Dell equipped with a Pentium 4 3.40 GHz processor and 512 MB of RAM. To conduct the PPMC for the discrepancy measures (described next), each cell at the large sample size took about 14.75 hours. Code for the MCMC and PPMC routines are available from the author upon request.

Discrepancy Measures

The choice of discrepancy measures should reflect (a) substantive aspects of the theory of interest and (b) features of the data that may not necessarily be adequately modeled. Three univariate functions are evaluated at the item level. Eight bivariate functions are evaluated for pairs of items.

Several of the discrepancy measures investigated involve univariate and bivariate observed and expected frequencies of response patterns for the (pair of) item(s) under consideration. The bivariate tables are two-way tables for the frequencies of values for two items given. For the pairing of items X_j and $X_{j'}$, the observed bivariate table is

		$X_{j'}$ = 1	$X_{j'}$ = 0
X_j	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

[6] I wish to thank Sandip Sinharay for sharing with me similar code which I subsequently adapted.

where $n_{kk'}$ is the number of examinees who have a value of k for item X_j and a value of k' for item $X_{j'}$. Similarly, the expected bivariate table is

		$X_{j'}$	
		1	0
X_j	1	$E(n_{11})$	$E(n_{10})$
	0	$E(n_{01})$	$E(n_{00})$

The elements of the expected table are the frequencies as implied by the IRT model, which may be obtained by integration over the distribution of θ_i (Chen & Thissen, 1997) which in the Bayesian framework is the posterior distribution.

In this study the approach adopted to obtain the expected frequencies is based on the conditional independence assumptions. For each subject, the assumption of local independence implies that the joint distribution of responses (conditional on the latent variable) to two items factors into the product of distributions of responses to each item separately (conditional on the latent variable):

$$P\left(X_{ij} = x_{ij}, X_{ij'} = x_{ij'} \mid \theta_i, \boldsymbol{\omega}_j, \boldsymbol{\omega}_{j'}\right) = P\left(X_{ij} = x_{ij} \mid \theta_i, \boldsymbol{\omega}_j\right) \times P\left(X_{ij'} = x_{ij'} \mid \theta_i, \boldsymbol{\omega}_{j'}\right). \quad (7)$$

The assumption of respondent independence implies that the elements in the expected table may be formed by summing the corresponding instantiations of Equation (7) over subjects. For example, the expected number of subjects responding to both items correctly is

$$E(n_{11}) = \sum_{i=1}^{n} P\left(X_{ij} = 1 \mid \theta_i, \boldsymbol{\omega}_j\right) \times P\left(X_{ij'} = 1 \mid \theta_i, \boldsymbol{\omega}_{j'}\right).$$

Analogous equations may be used to obtain expected counts for the remaining cells. In the simpler case of the univariate tables, a similar derivation allows for the calculation of

the expected values. Equivalently, the univariate values may be calculated by collapsing the two-way tables above. Appendix B provides a more complete description of this approach in comparison to alternative approaches.

Univariate Discrepancy Measures

The first discrepancy measure is the proportion correct, which is the sample mean. For item X_j:

$$PC_j = \frac{\sum_{i=1}^{N} X_{ij}}{N}. \tag{8}$$

Sinharay et al. (in press) reported the proportion correct to be an ineffective tool for conducting PPMC; it is hypothesized that the proportion correct will not prove useful in the current work.

The two other univariate discrepancy measures are X^2 and G^2 discrepancy measures for items (e.g., Fu et al., 2005), given respectively by

$$X_j^2 = \sum_{k=0}^{1} \frac{\left(n_k - E\left(n_k\right)\right)^2}{E\left(n_k\right)}; \tag{9}$$

$$G_j^2 = -2\sum_{k=0}^{1} n_k \ln \frac{E\left(n_k\right)}{n_k}. \tag{10}$$

Fu et al. (2005) found that, as a group, univariate discrepancy measures were less successful than bivariate discrepancy measures for detecting model inadequacies related to those studied here.

Bivariate Discrepancy Measures

X^2 and G^2 discrepancy measures for item-pairs (e.g., Chen & Thissen, 1997; Fu et al., 2005) are given, respectively, by

$$X^2_{jj'} = \sum_{k=0}^{1}\sum_{k'=0}^{1} \frac{\left(n_{kk'} - E\left(n_{kk'}\right)\right)^2}{E\left(n_{kk'}\right)}; \tag{11}$$

$$G^2_{jj'} = -2\sum_{k=0}^{1}\sum_{k'=0}^{1} n_{kk'} \ln \frac{E\left(n_{kk'}\right)}{n_{kk'}}. \tag{12}$$

Several correlational measures are explored. The covariance is given by

$$COV_{jj'} \frac{\sum_{i=1}^{N}\left(X_{ij} - \overline{X}_j\right)\left(X_{ij'} - \overline{X}_{j'}\right)}{N} = \frac{\left(n_{11}\right)\left(n_{00}\right) - \left(n_{10}\right)\left(n_{01}\right)}{N^2}. \tag{13}$$

The next discrepancy measures computed are Q_3 (Yen, 1984; 1993)

$$Q_{3_{jj'}} = r_{e_{ij}e_{ij'}} \quad \text{where} \quad e_{ij} = X_{ij} - E\left(X_{ij}\right), \tag{14}$$

and the closely related model-based covariance recommended for IRT model criticism by Reckase (1997a)

$$MODCOV_{jj'} = \frac{\sum_{i=1}^{N}\left(X_{ij} - E\left(X_{ij}\right)\right)\left(X_{ij'} - E\left(X_{ij'}\right)\right)}{N}, \tag{15}$$

where in the case of dichotomous observables $E\left(X_{ij}\right) = P\left(X_{ij} = 1 \mid \theta_i, \boldsymbol{\omega}_j\right)$ and is given by the IRF (Equation (1)).

McDonald and Mok (1995, in a frequentist framework) and Fu et al. (2005, in a PPMC framework) recommended the use of residual item covariances. A discrepancy measure that compares the covariance in the data to that based on the expected frequencies is given by

$$RESIDCOV_{jj'} = \frac{\left[\left(n_{11}\right)\left(n_{00}\right) - \left(n_{10}\right)\left(n_{01}\right)\right]}{N^2} - \frac{\left[E\left(n_{11}\right)E\left(n_{00}\right) - E\left(n_{10}\right)E\left(n_{01}\right)\right]}{E\left(N^2\right)}. \tag{16}$$

Correlational measures for linear association may be inappropriate in the case of dichotomous observables due to the nonlinearity of the associations and the dependence

on the mean (Mislevy, 1986b). Unlike the correlations, the odds ratio (and the log transform of it) is a measure of association that is not dependent on the marginal distributions of the observables (Liebetrau, 1983). We include as a discrepancy measure the odds ratio on the (natural) log scale (Agresti, 2002),

$$\ln(OR_{jj'}) = \ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right] = \ln(n_{11}) + \ln(n_{00}) - \ln(n_{10}) - \ln(n_{01}). \quad (17)$$

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. Principally related to the current work, they found the odds ratio to be an effective tool for detecting data-model misfit in several types of model misspecifications that induce local dependencies among the items. Note that because the log is a monotonic transformation, the PPP-value based on the log odds ratio will be identical to that based on the odds ratio.

The final discrepancy measure is the standardized log odds ratio residual for local dependence (Chen & Thissen, 1997):

$$STDLN(OR_{jj'}) = \frac{\ln\left[\frac{(n_{11})(n_{00})}{(n_{10})(n_{01})}\right] - \ln\left[\frac{E(n_{11})E(n_{00})}{E(n_{10})E(n_{01})}\right]}{\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}}. \quad (18)$$

HYPOTHESES

Based on the analysis of multidimensionality from a conditional covariance perspective and the performance of these discrepancy measures in other research, the following hypotheses were formed regarding the manipulated factors and the discrepancy measures.

H1: As the strength of the dependence on the auxiliary dimensions (a_{j2} and a_{j3} in the compensatory model and b_{j2} and b_{j3} in the conjunctive model) increases, it will become *easier* to detect data-model misfit in terms of pairs of items that reflect the same multiple dimensions.

H2: As the correlations among the dimensions increases, it will become *harder* to detect data-model misfit in terms of all types of item-pairs.

H3: As the proportion of items reflecting an auxiliary dimension increases, it will become *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 item-pairs that reflect the primary dimension only.

H4: As sample size increases, it will become *easier* to detect model misfit.

H5: The proportion correct statistic (Equation (8)) will not be an effective quantity for detecting multidimensionality.

H6: As a group, the bivariate discrepancy measures (Equations (11) – (18)) will be more effective then the univariate discrepancy measures (Equation (8) – (10)).

H7: The most effective quantities will be the residual covariance (Equation (16)), log odds ratio (Equation (17)), and the standardized log odds ratio difference (Equation (18)).

RESULTS

Results will be presented separately for the different data-generating models (unidimensional, compensatory, conjunctive). Several patterns emerged; where warranted, the presentation of redundant patterns is omitted. For the multidimensional data-generating models, the results for the large sample size will be 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. The effects of sample size will be incorporated into the presentation of the proportion of extreme PPP-values.[7]

Unidimensional Data

There are three null conditions in which the data are unidimensional, corresponding to the three sample sizes. Within each condition, the analysis is replicated 50 times. For any replication, each univariate discrepancy measures is evaluated 32 times (once for each item) leading to 32 PPP-values. Each bivariate discrepancy measure is evaluated 496 times (one for each unique pairing of items) leading to 496 PPP-values. The PPP-values from the 32 items (likewise, 496 item-pairs) for each univariate (bivariate) discrepancy measure are pooled. This pooling follows from an exchangeability assumption (de Finetti, 1964). Given that the data were generated from a unidimensional model, all items (item-pairs) have the same dimensional structure, namely, they all reflect the same single latent dimension. In addition to the pooling of

[7] In some cases the G^2 measure for item-pairs, log odds ratio, and standardized log odds ratio residual (equations (12), (17), and (18)) could not be computed due to zero frequencies for counts of bivariate response patterns (Chen & Thissen, 1997). This was quite rare. In the null conditions, for the sample sizes of 250 and 750, PPP-values could not be computed for these measures in 0.13% and 0.01% of the cases, respectively. In the non-null conditions, PPP-values could not be computed for these measures in 1.90%, 0.13%, and 0.004% of the cases for the sample sizes of 250, 750, and 2500, respectively. These cases were ignored from the analyses.

items or item-pairs for each replication, results from all 50 replications in each condition are pooled.

Figure 12 contains eleven panels, one for each discrepancy measure. In each panel there are line plots for the distributions of the PPP-values. The horizontal axis spans the full range of the PPP-values from 0 to 1 (the left and right endpoints within each panel). The heights of the points are proportional to the relative frequency of the PPP-values at or near that value, akin to a histogram. For each discrepancy measure, the line plots are virtually indistinguishable, indicating that the distributions are similar across the different sample sizes.

As observed in Figure 12, all the distributions of PPP-values are symmetric around .5. As a consequence, the median PPP-values hover around .5. This is depicted explicitly in Figure 13, where the median PPP-value for the discrepancy measure (based on 50 replications of each condition and pooling all items or item-pairs) at each sample size is plotted. Note that the horizontal axis (sample size) is not drawn to scale. The lines connecting the points in each panel are included as a visual tool and are not meant to represent any function for interpolation.

Figure 12: Distributions of PPP-values for 11 discrepancy measures based on unidimensional data.

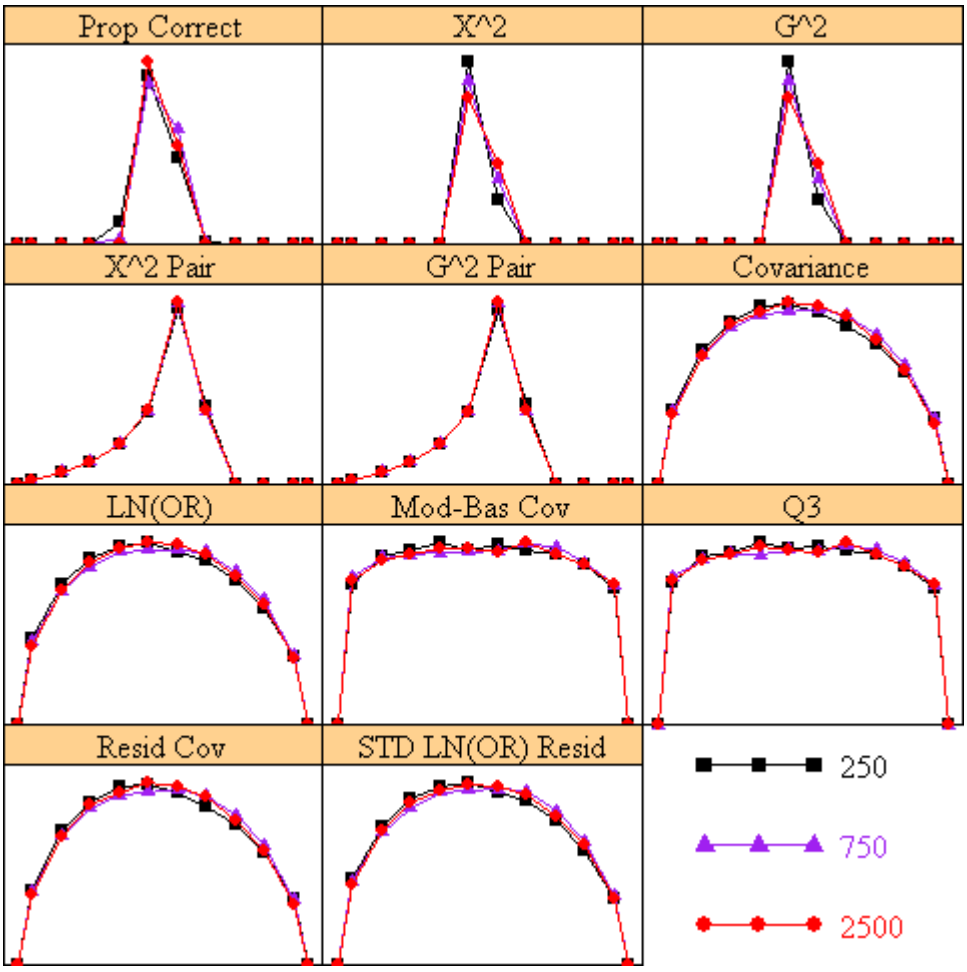
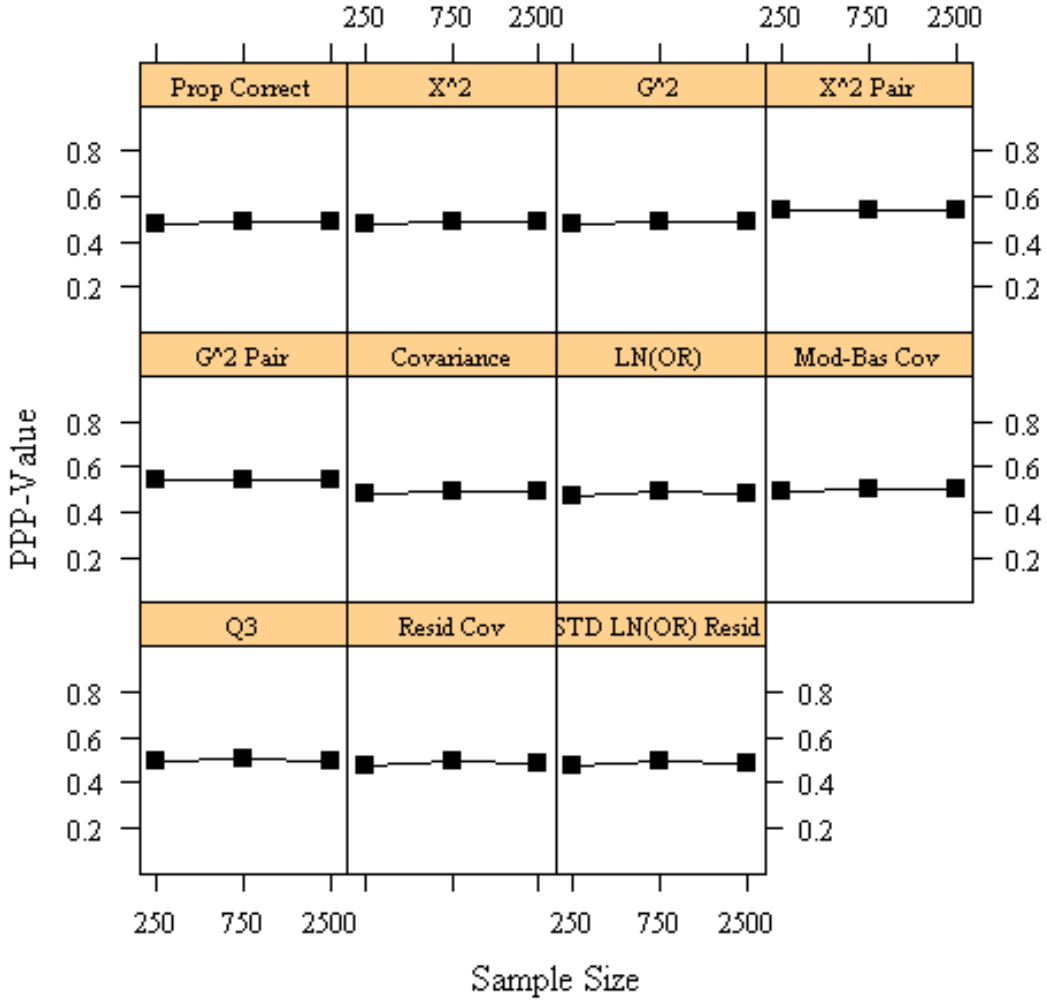


Figure 13: Median PPP-values for 11 discrepancy measures based on unidimensional data.



The finding of median PPP-values near .5 is consistent with theoretical results on the behavior of PPP-values under null conditions (Bayarri & Berger, 2000; Meng, 1994; Robins et al., 2000). A PPP-value around .5 indicates that the realized value of the discrepancy measure falls in the middle of the posterior predictive distribution. Substantively, this indicates solid data-model fit, as would be expected under null conditions, as the estimated model and the true data generating process are the same.

The discrepancy measures differ in their distributions of PPP-values in terms of variability. The proportion correct, X^2, and G^2 measures are tightly distributed around .5. The X^2 and G^2 measures for item-pairs exhibit slightly more variation. The covariance, log odds ratio, residual covariance, and standardized log odds ratio residual exhibit more variability. Finally, the model-based covariance and Q_3 exhibit the most variability and are close to being uniformly distributed.

As discussed previously, though PPP-values are distributed around .5 under null conditions, they are not necessarily uniformly distributed, even asymptotically (Bayarri & Berger, 2000; Robins et al., 2000). The result is that employing PPP-values to conduct hypothesis testing leads to conservative inferences (Fu et al., 2005; Sinharay et al., in press; Sinharay & Stern, 2003).

In the current work, this conservativeness is evident in Table 2, which contains the proportions of extreme PPP-values for the discrepancy measures for each sample size. Extreme PPP-values are defined as those below .05 or above .95. In a hypothesis testing framework, the contents of Table 2 are Type I error rates using .05 and .95 as critical values (i.e., in a two-tailed test with α = .10). All the values in Table 1 are below .10, indicating that use of PPP-values in hypothesis testing leads to a conservative test. Note however, that the model-based covariance and Q_3 exhibit empirical Type I error rates close to the nominal rate of .10.

Table 2: Proportion of replications with extreme PPP-values (i.e., PPP-value < .05 or >.95) when the data follow a unidimensional model.

		Discrepancy Measure									
Sample Size	Prop Correct	X^2	G^2	X^2 Pair	G^2 Pair	Cov	LN(OR)	Mod-Based Cov	Q_3	Resid Cov	Std LN(OR) Resid
250	.00	.00	.00	.00	.01	.04	.04	.08	.08	.04	.04
750	.00	.00	.00	.00	.00	.04	.04	.09	.09	.04	.04
2500	.00	.00	.00	.00	.00	.03	.04	.08	.08	.03	.04

As discussed previously we follow Gelman et al. (1996) and Stern (2000) in advocating the use of PPMC as a model diagnostic rather than as a test, which is consistent with Bayesian and frequentist approaches to local dependence assessment which cautions the analyst from relying solely on statistical results when evaluating model fit (Chen & Thissen, 1997; Sinharay, 2005; Zenisky et al., 2003). Nevertheless, uniformity of PPP-values under null conditions may be desirable (Berkhof et al., 2004). With this in mind, Figure 12 and Table 2 suggest that (a) the proportion correct, X^2, and G^2 measures perform similarly and are the worst, (b) the X^2 and G^2 measures for pairs of items perform similarly to each other and are slightly better than their univariate counterparts, (c) the covariance, log odds ratio, residual covariance, and standardized log odds ratio perform similarly to one another and are better than all the previously listed measures, and (d) the model-based covariance and Q_3 are comparable, better than all the other measures, and close to optimal.

These results constitute a reason to prefer the model-based covariance and Q_3 above the other discrepancy measures. After presenting the results for the multidimensional data, this is discussed in a broader context of evaluating the discrepancy measures under null and non-null conditions.

Compensatory Multidimensional Data

Results of PPMC when the data are multidimensional will be presented for the large sample size first. The behavior of each discrepancy measure across the other levels of the manipulated factors (strength of dependence, correlations among dimensions, and proportion of items reflecting multiple dimensions) will be presented. As above, PPP-values obtained from calculating the discrepancy measure on multiple items (item-pairs) are pooled following exchangeability assumptions, as well as across the 50 replications within conditions. As described in more detail below (see also Appendix C), which PPP-values are pooled will vary depending on (a) which items reflect multiple dimensions and (b) whether the discrepancy measure is univariate or bivariate.

As implied by theory (Meng, 1994), Figure 12 suggests that under null conditions, PPP-values vary around .5. This is true regardless of the sample size or discrepancy measure under consideration. Accordingly, PPP-values near .5 are indicative of adequate data-model fit. PPP-values that deviate from .5 are suggestive of a lack of data-model fit. In discussing the behavior of the discrepancy measures under the non-null conditions of multidimensionality, PPP-values will be interpreted accordingly. More specifically, PPP-values near .5 indicate that the PPMC has *failed* to detect the multidimensionality as PPP-values near .5 are wholly consistent with the hypothesis that model fits the data well. PPP-values farther away from .5 indicate that the PPMC has *succeeded* (to some degree)

in detecting the multidimensionality, as such PPP-values are inconsistent with the hypothesis that the model fits the data well. Success is not an all-or-none evaluation. As discussed below, much can be gained from investigating the relative values of the different discrepancy measures under the studied conditions.

Median PPP-values for 2500 Examinees

Univariate Discrepancy Measures

Figure 14 plots median PPP-values for the proportion correct discrepancy measure. There are 16 panels in the plot, corresponding to the combinations of the four levels of strength of dependence with the four levels of the correlations among the latent variables. The four rows (of four plots) correspond to the different values of the strength of dependence on the auxiliary dimensions. Proceeding from top to bottom, the strength of dependence increases (a_{j2}, a_{j3} = 0.25, 0.50, 0.75, 1.00). The four columns (of four plots) correspond to the different values of the correlations between the dimensions. Proceeding from left to right, the correlations increase (ρ_{21}, ρ_{31}, ρ_{32} = 0.0, 0.3, 0.7, 0.9). For example, the plot in the second row, third column corresponds to the case where the strength of dependence (a_{j2}, a_{j3}) is 0.50 and the correlations between the latent dimensions (ρ_{21}, ρ_{31}, ρ_{32}) are .7.

Within each panel, the vertical axis is the PPP-value and the horizontal axis is the proportion of items influenced by the second or third dimension. Moving left to right, the three points on the horizontal axis are the low, medium, and high proportion of multidimensional items (four, 16, and 28 items, respectively). There are two sets of points within each panel. Each black square (one at each position on the horizontal axis) is a median PPP-value for items that only reflect the first dimension. In the key, this is

denoted as '(1)' to indicate that the black squares represent items that only reflect the first dimension. Each red circle is a median PPP-value for items that reflect multiple dimensions. That is, all items that reflect the first and second dimension were pooled with items that reflect the first and third dimension.[8] In the key, this is denoted as '(1,2) (1,3)' to indicate that the red circles represent items that reflect the first and second or first and third dimensions. As in the previous figure, lines connecting points are plotted for visual ease.

The lower left panel is the condition in which the strength of dependence is strongest (a_{j2}, a_{j3} = 1.0) and the correlations among the dimensions are the smallest (ρ_{21}, ρ_{31}, ρ_{32} = 0.0). This represents the combination of these factors that was hypothesized to be the easiest condition in which to detect the multidimensionality. The red circles lie almost exactly on top of the black squares. For this condition, the proportion correct is unable to distinguish between items that reflect multiple dimensions and items that only reflect the primary dimension, regardless of the number of items that reflect multiple dimensions. This pattern holds across all 16 conditions in Figure 14.

Figures 15 and 16 plot the median PPP-values for the univariate X^2 and G^2 discrepancy measures, respectively, following the structure in Figure 14. In both plots, in all conditions, the red circles closely mirror the black squares.

[8] See Appendix C for an explanation and justification of the assumptions underlying the choices regarding what to pool.

Figure 14: Median PPP-values for the proportion correct when the data follow a compensatory MIRT model and N=2500.

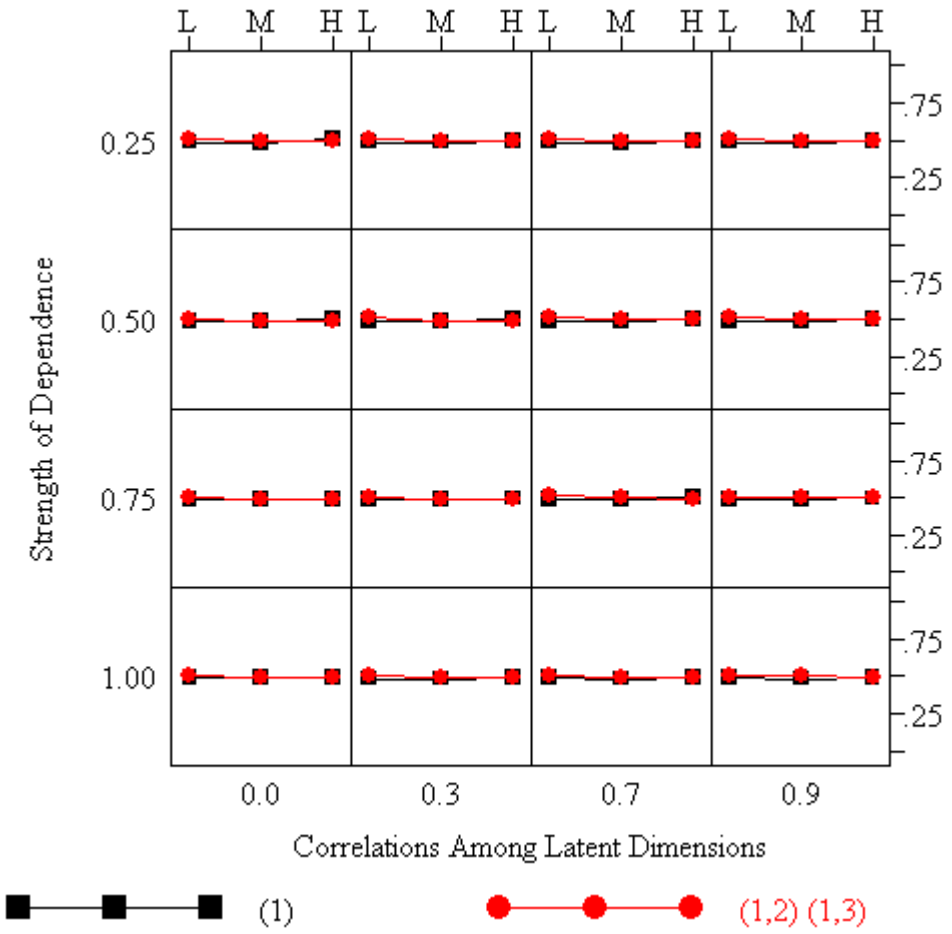


Figure 15: Median PPP-values for X^2 when the data follow a compensatory MIRT model and N=2500.

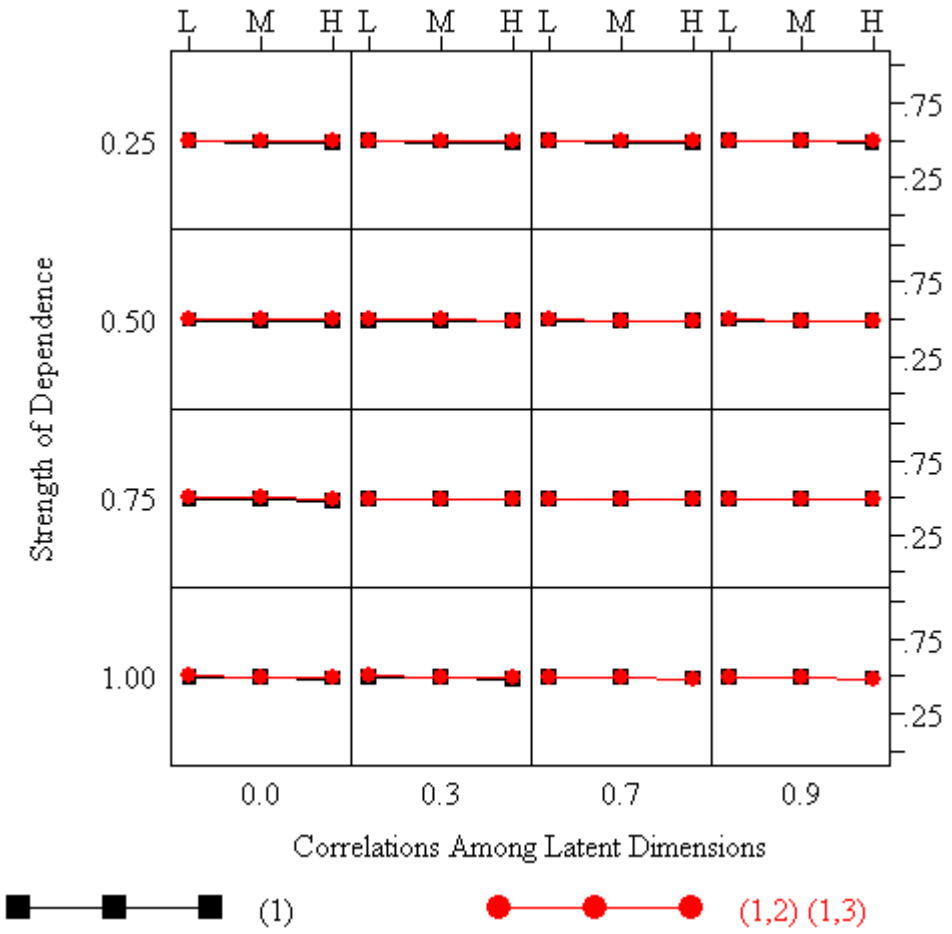
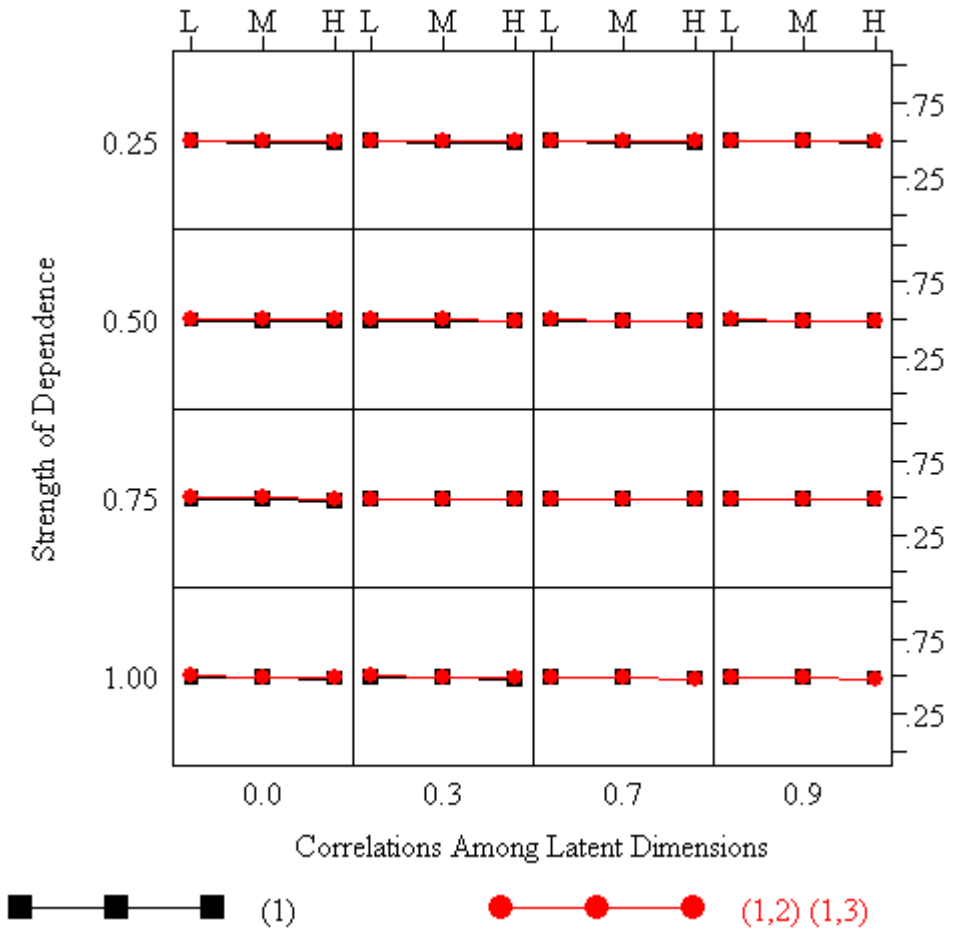


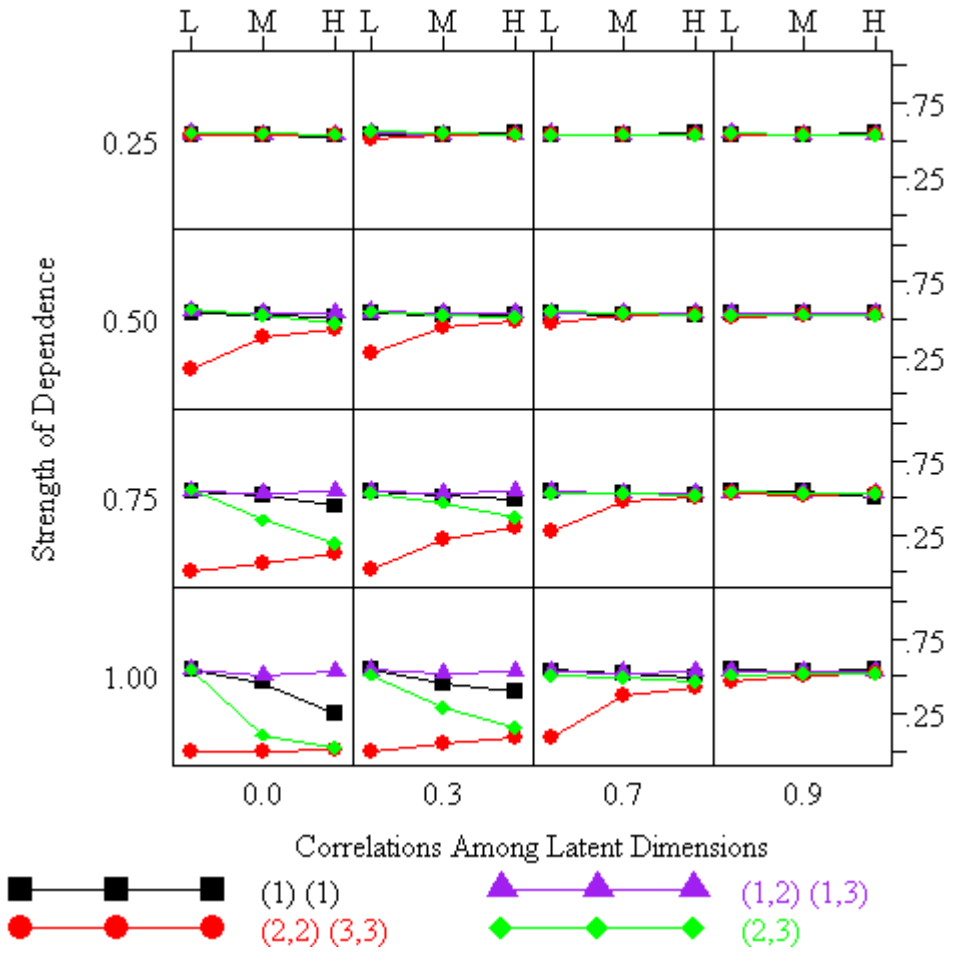
Figure 16: Median PPP-values for G^2 when the data follow a compensatory MIRT model and N=2500.



Bivariate Discrepancy Measures

Figure 17 plots the median PPP-values for the X^2 discrepancy measure for pairs of items. The structure of the panels follows those in Figures 14-16 in which the 16 panels correspond to the combinations of the four levels of strength of dependence with the four levels of the correlations among the latent variables. Likewise, within each panel, the three points on horizontal axis correspond to low, medium, and high proportion of items that reflect multiple dimensions.

Figure 17: Median PPP-values for X^2 for item-pairs when the data follow a compensatory MIRT model and N=2500.



The contents of each panel are more complex, owing to the discrepancy measure being a bivariate function evaluated for pairs of items. With three latent dimensions underlying the items, there are four distinct types of item-pairs. The first type is the pairing of two items that reflect the first dimension only. These are plotted as black squares and are denoted as '(1),(1)' in the key.

The second type is the pairing of one item that only reflects the first dimension and another item that reflects multiple dimensions. These include pairings in which (a)

one item reflects only the first dimension and the second item reflects the first and second dimension, and (b) one item reflects only the first dimension and the second item reflects the first and third dimension. These item-pairs are plotted with purple triangles and are labeled as '(1,2) (1,3)' in the key.

The third type is the pairing of two items that reflect *the same* multiple dimensions. These include pairings in which both items reflect the first and second dimension and parings in which both items reflect the first and third dimensions. These item-pairs are plotted with red circles and are labeled as '(2,2) (3,3)' in the key.

The final type is the pairing of two items that reflect *different* multiple dimensions. More specifically, these are the pairings in which one item reflects the first and second dimension and the second item reflects the first and third dimension. This type of item-pair is plotted with green diamonds and labeled '(2,3)' in the key.

Item-pairs of each type were pooled, reflecting exchangeability assumptions. For example, item-pairs in which both items reflect the first and second dimension are pooled with item-pairs in which both items reflect the first and third dimension. See Appendix C for the foundation and an evaluation of these assumptions.

Figures 18 though 24 plot the median PPP-values for the remaining discrepancy measures. The results for the X^2 and G^2 discrepancy measures (Figures 17 and 18) are quite similar. The results for the other bivariate discrepancy measures (Figures 19-24) are also similar to each other.

Figure 18: Median PPP-values for G^2 for item-pairs when the data follow a compensatory MIRT model and N=2500.

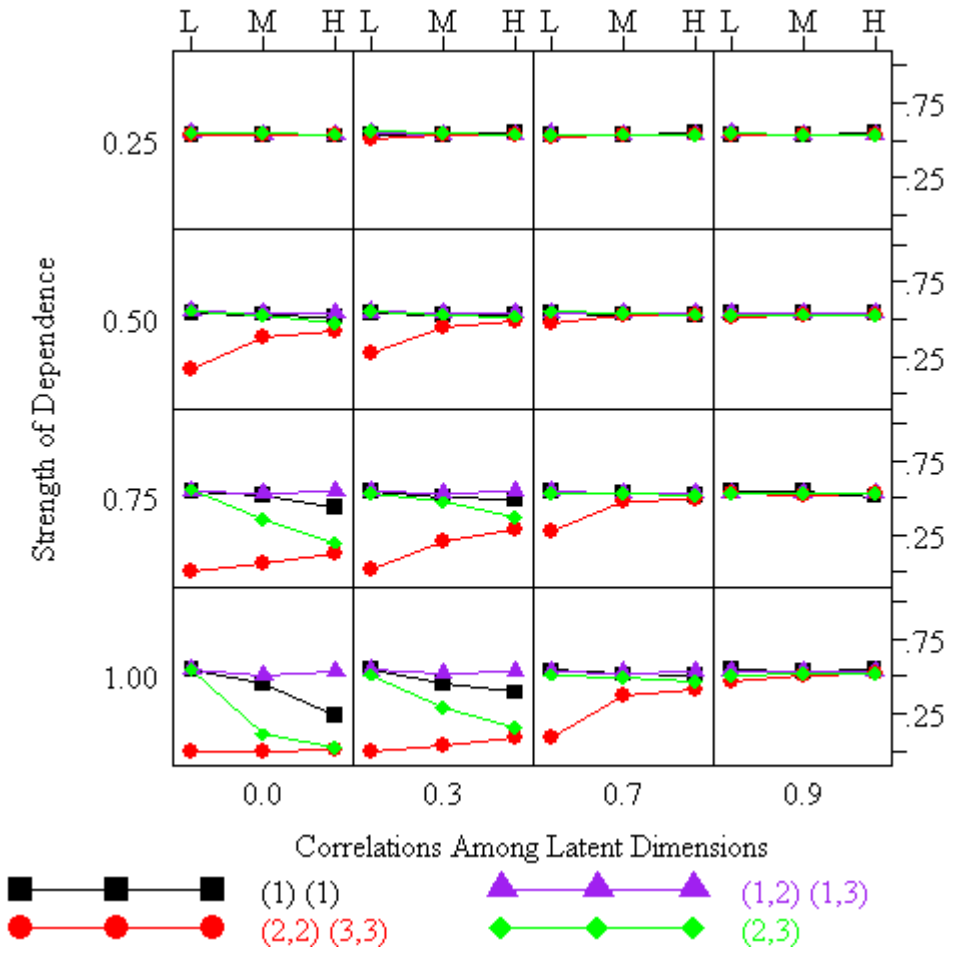


Figure 19: Median PPP-values for the covariance when the data follow a compensatory MIRT model and N=2500.

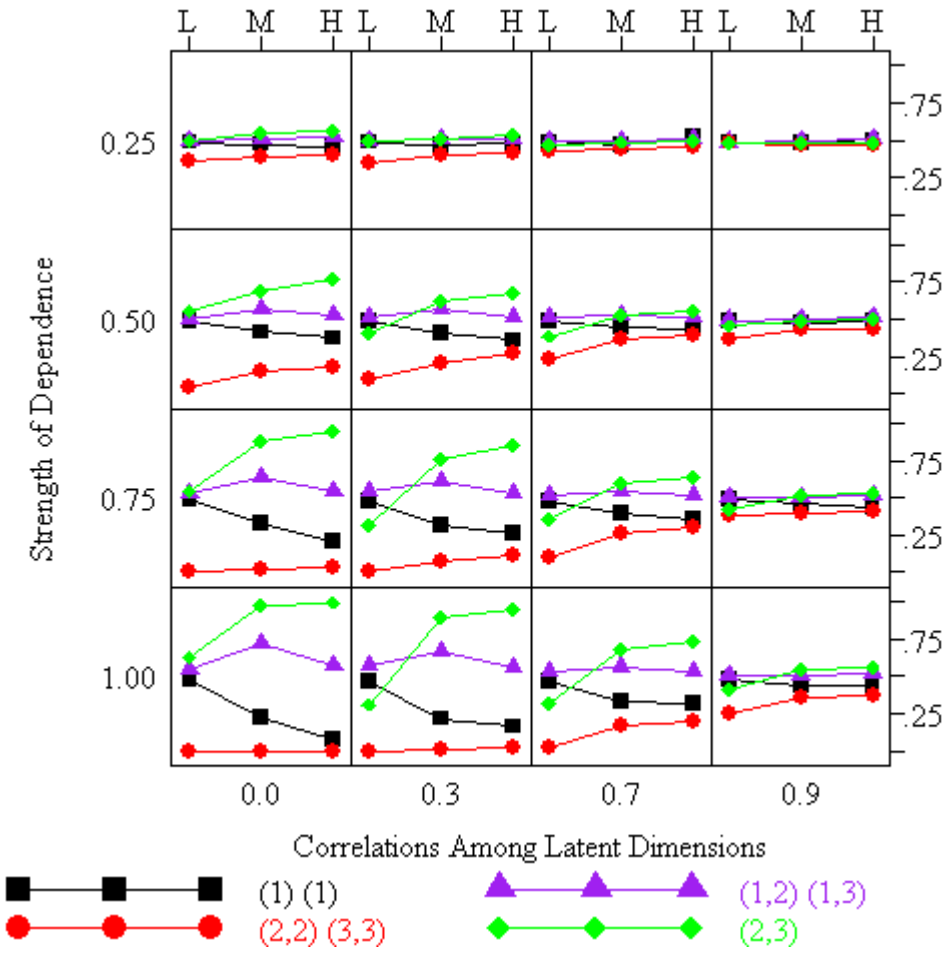


Figure 20: Median PPP-values for the log odds ratio for item-pairs when the data follow a compensatory MIRT model and N=2500.

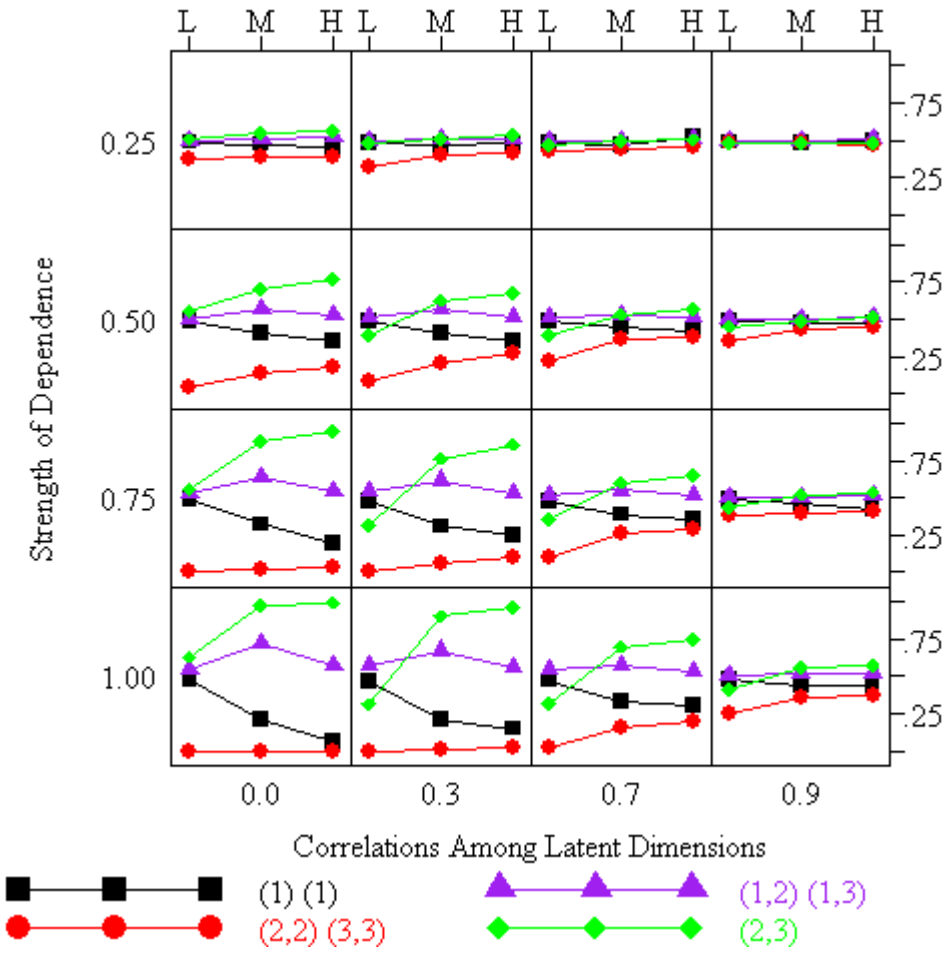


Figure 21: Median PPP-values for the model-based covariance for item-pairs when the data follow a compensatory MIRT model and N=2500.

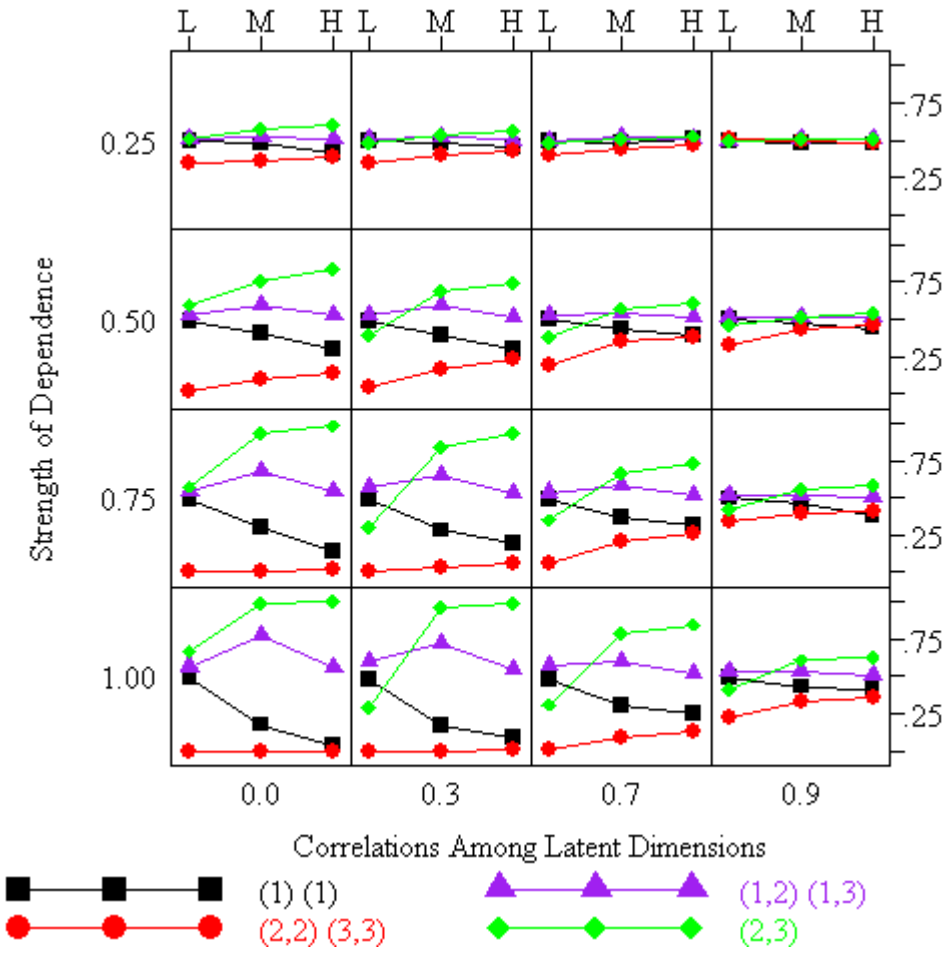


Figure 22: Median PPP-values for Q_3 for item-pairs when the data follow a compensatory MIRT model and N=2500.

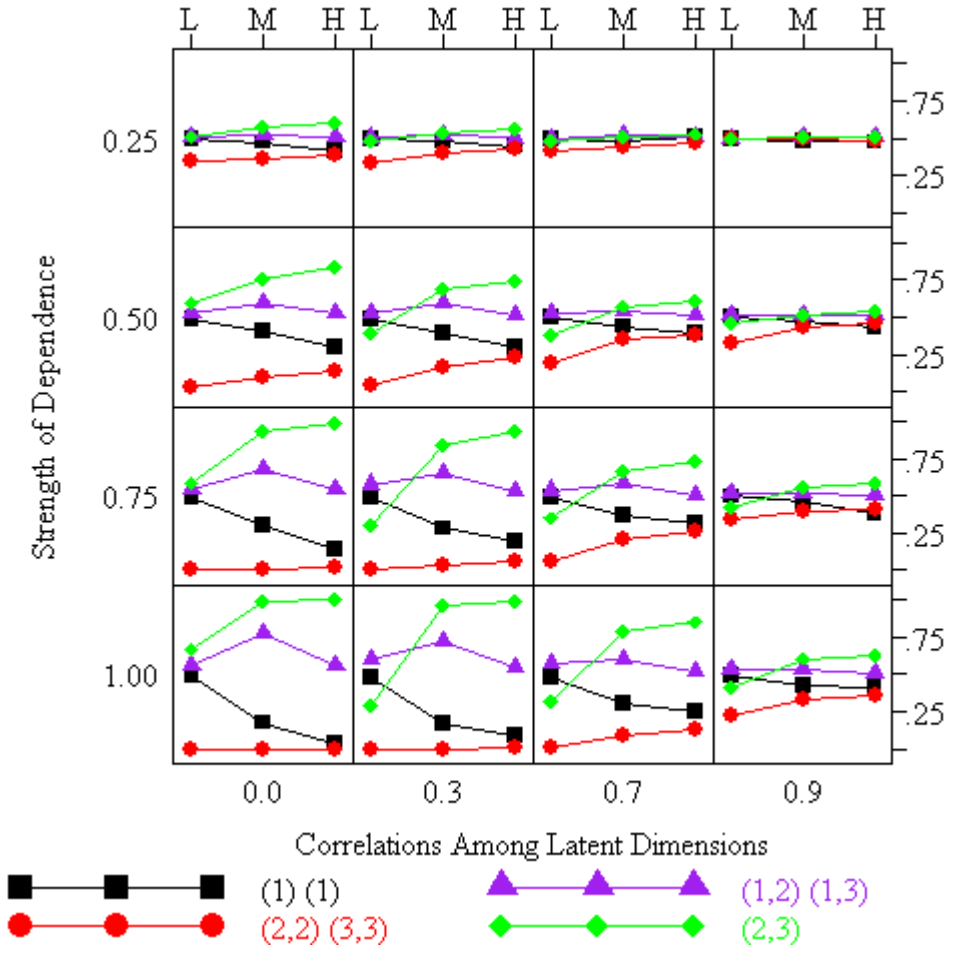


Figure 23: Median PPP-values for the residual covariance for item-pairs when the data follow a compensatory MIRT model and N=2500.

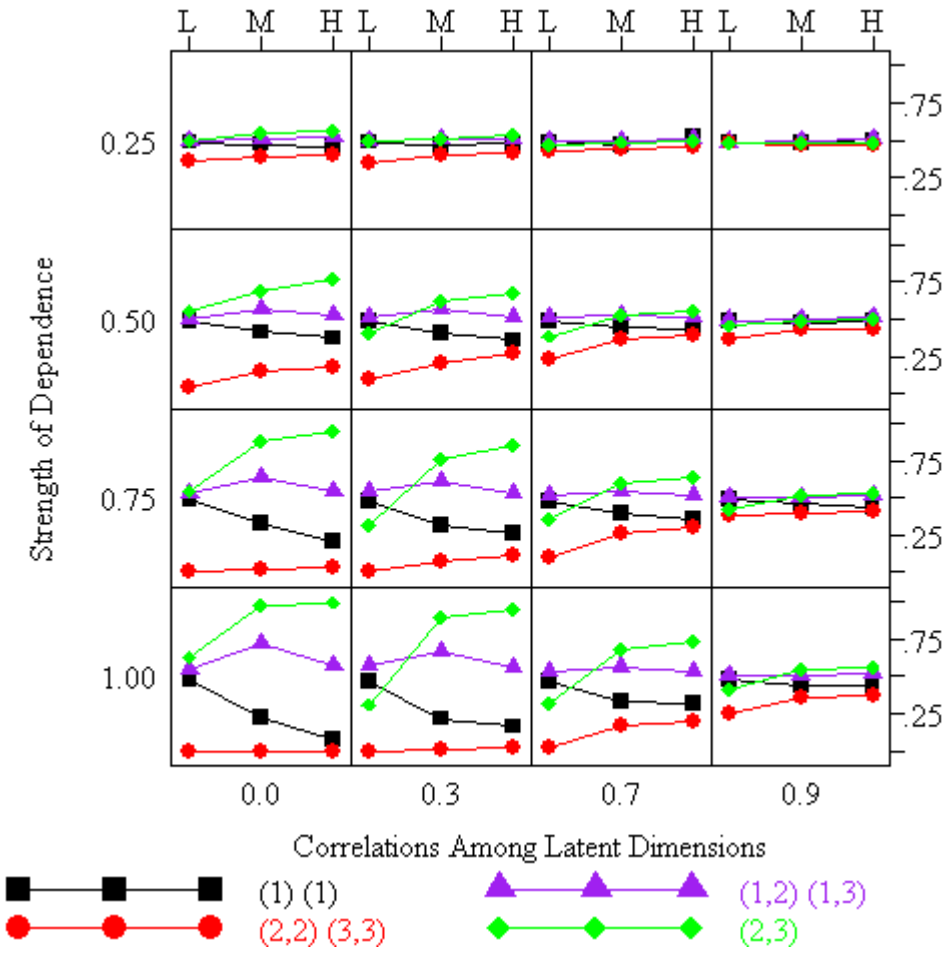
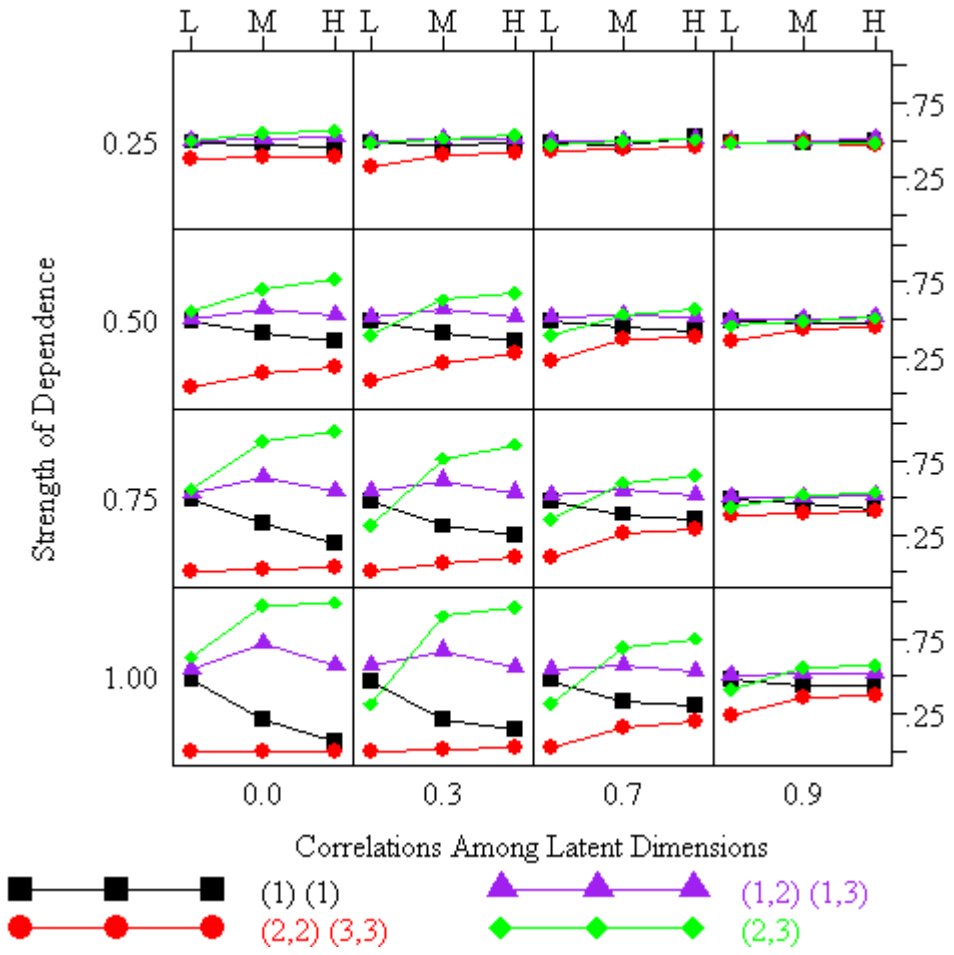


Figure 24: Median PPP-values for the standardized log odds ratio residual for item-pairs when the data follow a compensatory MIRT model and N=2500.



Proportion of Extreme PPP-values For 2500 Examinees

Though useful for tracking the general behavior the PPP-values across the conditions, the median is but one summary statistic, and is not optimal for summarizing the extreme values that lie in the tails. This was observed in the analysis of the PPMC for the null condition of unidimensional data. Though the medians for all discrepancy measures hovered around .5 (Figure 13), the proportion of extreme values varied considerably (Table 2), owing to the differences of the distributions (Figure 12).

This section describes the efficacy of the PPMC for detecting multidimensionality in terms of the propensity for the discrepancy measures to yield extreme PPP-values. The univariate discrepancy measures are not included as the median PPP-values indicated that they were not sensitive to any of the manipulated factors (Figures 14-16). As in the analysis of PPP-values under null conditions, 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 $\alpha = .10$.

Proportions of Extreme PPP-values by Type of Item-Pair

Table 3 presents the proportions of extreme PPP-values for the bivariate discrepancy measures for item-pairs in which both items reflect the same multiple dimensions. The first two columns list the levels of the strength of dependence in terms of the item parameters and the correlations between the dimensions, respectively. The relative performances of the different discrepancy measures are consistent across the different conditions. The proportions for X^2 and G^2 are quite close to one another (and are frequently equal) and are always the lowest, though in some conditions other measures have values as low as these. The proportions for the model-based covariance and Q_3 are quite close to each other (and are frequently equal). One or both of these measures is always the largest, though in some conditions other measures have values equal to them. The results for the covariance and the residual covariance are always the same, as are the results for the log odds ratio and the standardized log odds ratio residual. The results for the covariance (and the residual covariance) are similar to those of the log odds ratio (and the standardized log odds ratio), differing by no more than .04.

Table 3: Proportion of replications with extreme PPP-values (i.e., PPP-value < .05 or >.95) for item-pairs that reflect the same multiple dimensions when data follow a compensatory MIRT model, N=2500, and the proportion of items is low.

		Discrepancy Measure							
a_{j2}, a_{j3}	ρ	X^2 Pair	G^2 Pair	Cov	LN (OR)	Mod-Based Cov	Q_3	Resid Cov	Std LN(OR) Resid
.25	0.0	.01	.01	.05	.03	.10	.11	.05	.03
	0.3	.00	.00	.06	.07	.13	.13	.06	.07
	0.7	.01	.01	.05	.07	.15	.15	.05	.07
	0.9	.00	.00	.00	.01	.03	.03	.00	.01
.50	0.0	.18	.19	.57	.61	.69	.70	.57	.61
	0.3	.03	.03	.29	.33	.53	.53	.29	.33
	0.7	.03	.03	.14	.15	.20	.20	.14	.15
	0.9	.00	.00	.01	.03	.06	.06	.01	.03
.75	0.0	.93	.93	.98	.98	.98	.98	.98	.98
	0.3	.72	.73	.92	.91	.96	.96	.92	.92
	0.7	.09	.10	.30	.34	.46	.46	.30	.34
	0.9	.00	.00	.02	.04	.11	.11	.02	.04
1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	0.3	.99	.99	1.0	1.0	1.0	1.0	1.0	1.0
	0.7	.34	.35	.66	.70	.81	.81	.66	.70
	0.9	.01	.01	.08	.11	.18	.18	.08	.11

These patterns of similarities held for the remaining types of item-pairs in this condition, and in the remaining levels of sample size and the proportion of items reflecting multiple dimensions. That is, the proportions of extreme PPP-values for X^2 and G^2 are always quite close. The proportions for the covariance and the residual covariance are quite close (though not always equal), which in turn are also close to the log odds ratio and the standardized log odds ratio residual. Lastly, the proportions for the model-based covariance and Q_3 are quite close.

Tables for the remaining types of item-pairs and the remaining conditions will not be presented on space considerations. The results for selected discrepancy measures will be presented graphically here. Owing to the similarity of the results for some of the measures, presenting a few is sufficient to summarize the results of all. More specifically, the proportion of extreme PPP-values for X^2, the log odds ratio, and the model-based covariance will be presented and discussed in more detail. The results for X^2 are representative of G^2 also; the results for the model-based covariance are representative of Q_3 also. The results for the log odds ratio are representative of the covariance, residual covariance, and the standardized log odds ratio residual.

The panels in Figure 25 plot the proportion of extreme PPP-values for item-pairs that reflect the same multiple dimensions. The panels are structured similarly to those that display the median PPP-values across various conditions. The four rows of the panels correspond to the different levels of strength of dependence and the four columns correspond to the different levels of the correlations between the dimensions. Within each panel, the three points along the horizontal axis correspond to the low, medium, and high conditions for the proportion of multidimensional items. The plots in Figure 25

differ from the earlier plots in two important ways. The first is that the vertical axis is the proportion of times an extreme PPP-value was obtained. Second, the different points do not correspond to different types of item-pairs, but to different discrepancy measures, as indicated in the key.

Figure 25: 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 the data follow a compensatory MIRT model and N=2500.

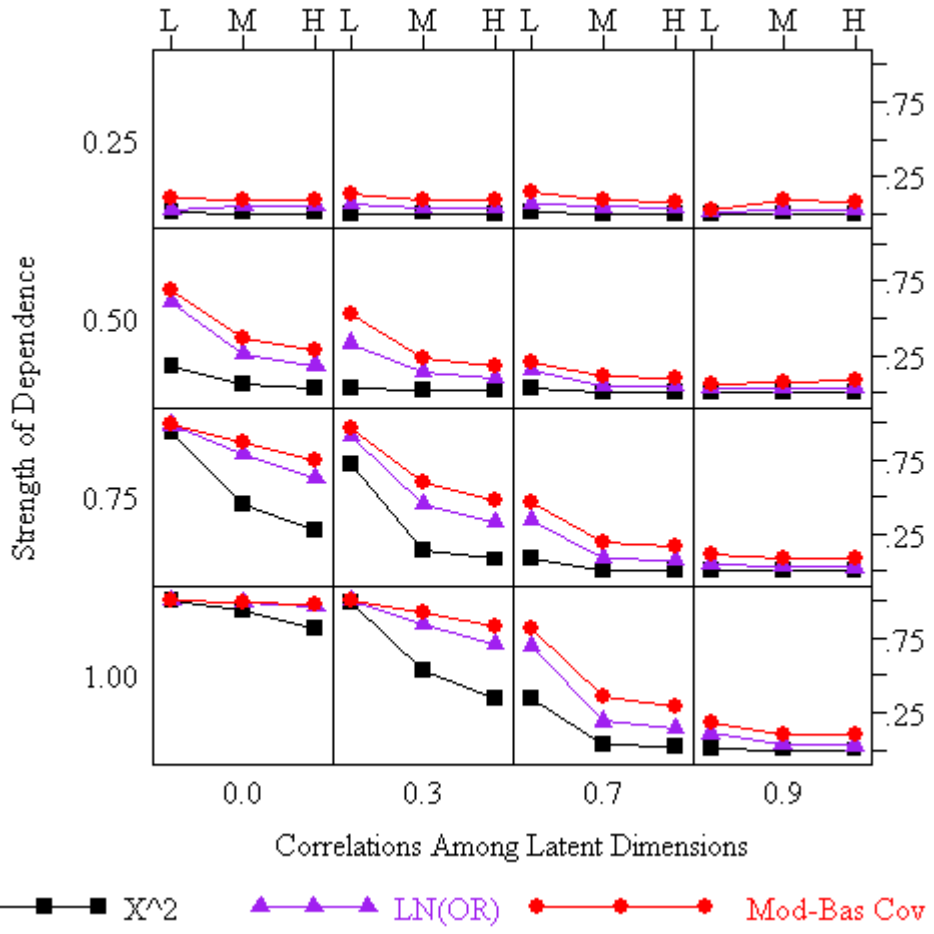


Figure 26 presents the results for item-pairs that reflect different multiple dimensions. Figure 27 presents the results for item-pairs in which both items reflect the

primary dimension only. 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 did not meaningfully deviate from .5 for any of the discrepancy measures. Suffice it to say the proportions of extreme PPP-values are quite low and do not show a systematic pattern.

Figure 26: 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 the data follow a compensatory MIRT model and N=2500.

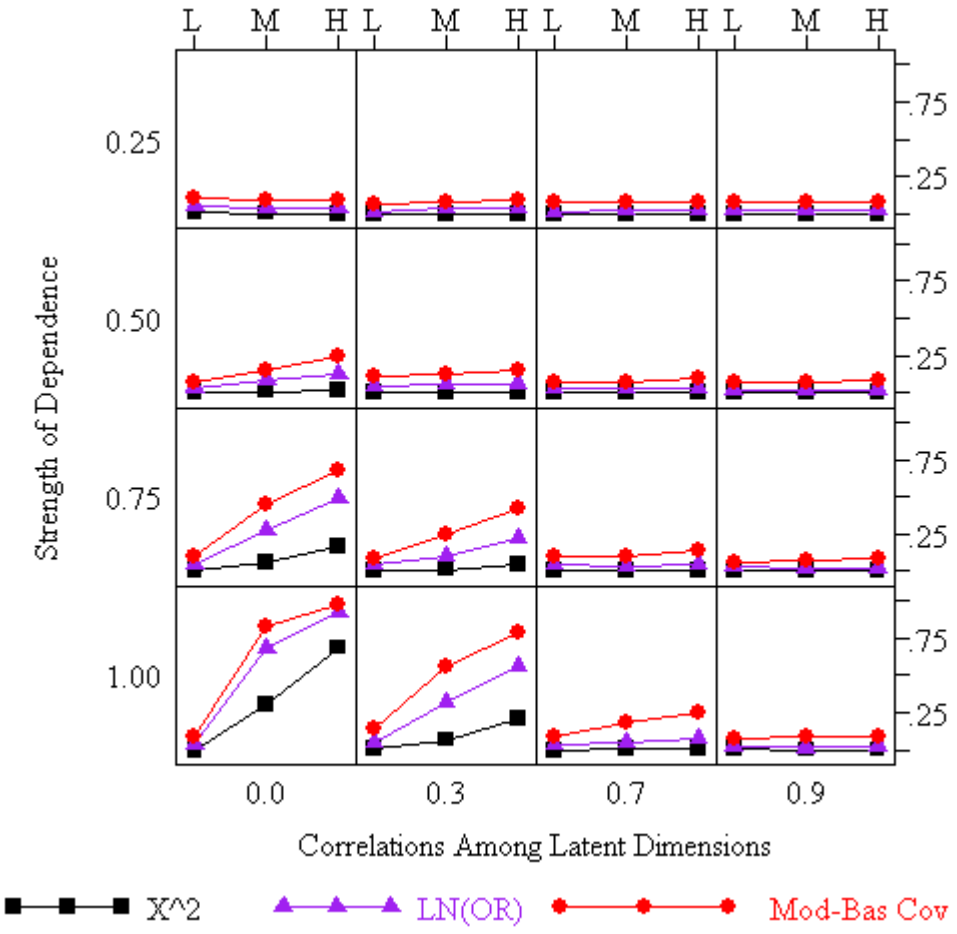
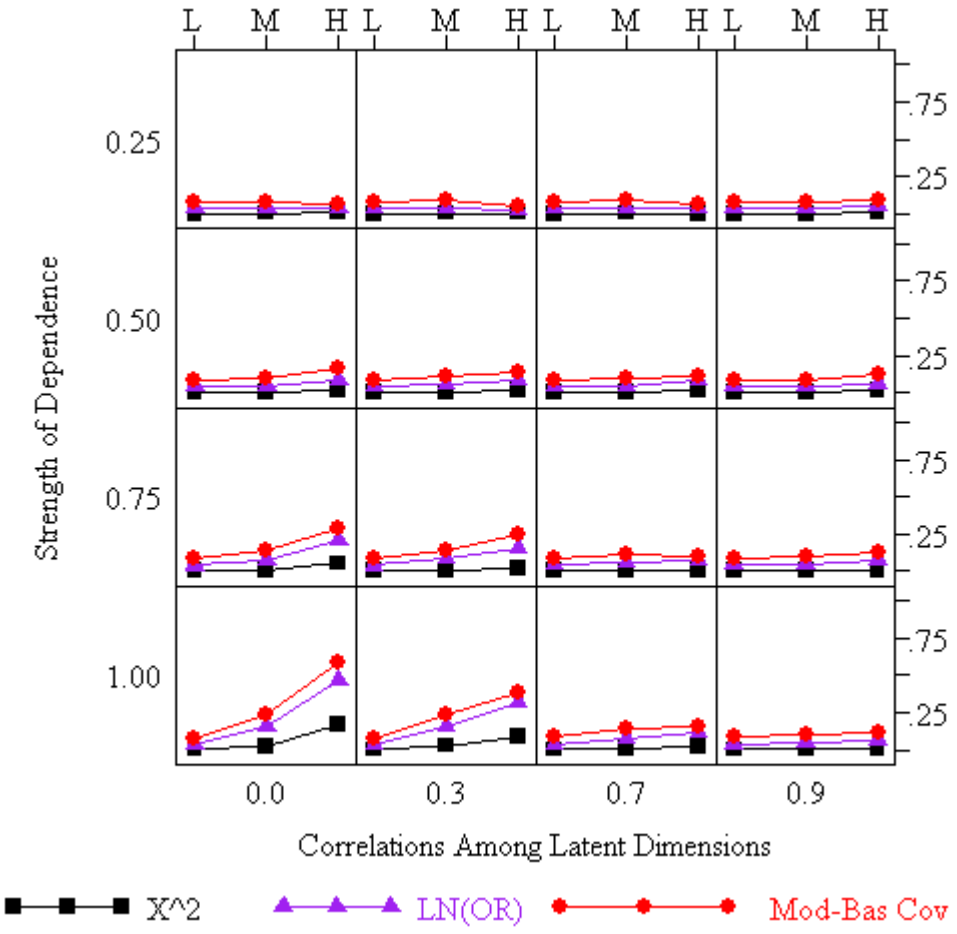


Figure 27: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for select discrepancy measures for item-pairs that reflect the primary dimension only when the data follow a compensatory MIRT model and N=2500.



Proportions of Extreme PPP-values by Sample Size

We turn our attention to the influence of sample size on PPMC. For ease of exposition, only the model-based covariance will be presented; recall that the model-based covariance is also representative of Q_3. Figures 28-30 display plots of the proportions of extreme PPP-values for item-pairs that reflect (a) the same multiple dimensions, (b) different multiple dimensions, and (c) the primary dimension only, respectively, at each condition. The panels are similarly structured to those presented

earlier. The sixteen panels correspond to the combinations of strength of dependence and the correlation among the dimensions.

Within each panel, the vertical axis is the proportion of extreme PPP-values. The three sets of points (and lines) correspond to the three proportions of multidimensional items: low (black squares), medium (purple triangles), and high (red circles). The horizontal axis corresponds to sample size. Thus the trend for a set of points within a panel corresponds to the power as sample size increases from 250 to 750 and to 2500.

Figure 28: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect the same multiple dimensions when the data follow a compensatory MIRT model.

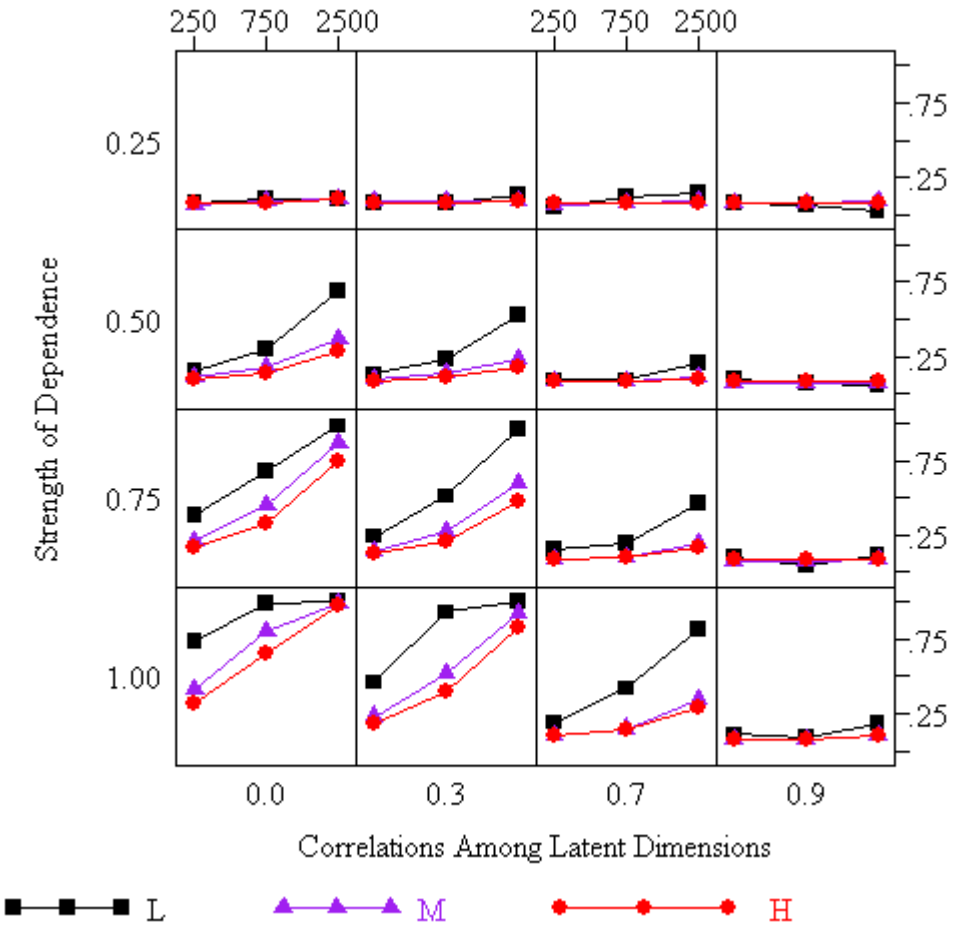


Figure 29: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect different multiple dimensions when the data follow a compensatory MIRT model.

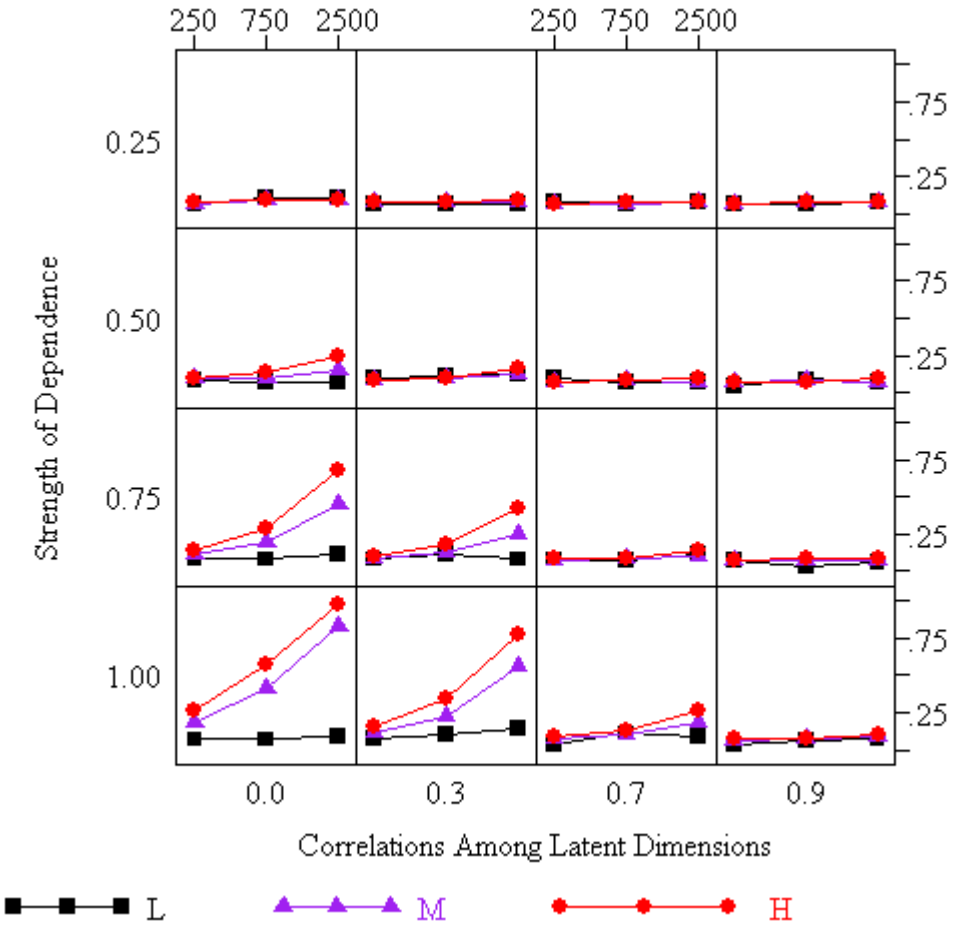
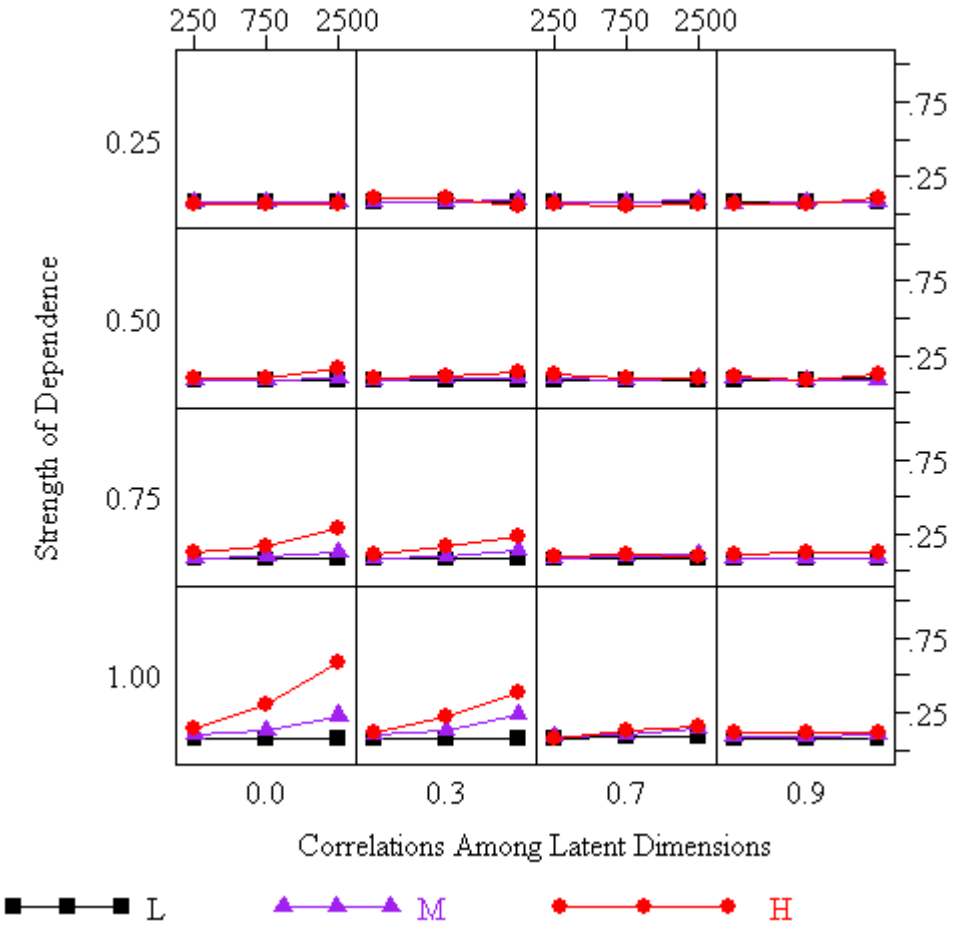


Figure 30: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect the primary dimension only when the data follow a compensatory MIRT Model.



Discussion

Univariate Discrepancy Measures

Beginning with the univariate discrepancy measures (Figures 14-16), it is observed that for all levels of the strength of dependence and all levels of the correlations between the dimensions, the circles are essentially at the same points as the squares. Substantively, the proportion correct, X^2, and G^2 discrepancy measures were unable to detect the multidimensionality in any of the conditions. This behavior was hypothesized

for the proportion correct and is consistent with findings in other applications of PPMC (Sinharay et al., in press). Expanding on a conjecture by Sinharay et al. (in press) a possible explanation for its ineffectiveness in IRT models is that the inclusion of a unique location parameter in the 2-PL is sufficient to recover the proportion correct for the item, regardless of how misspecified the model is relative to the data.

The inability of the X^2 and G^2 discrepancy measures to detect the multidimensionality is consistent with hypothesis H6 that univariate discrepancy measures will not perform as well as bivariate measures. When multidimensionality goes unaccounted for there are sources of covariation that induce associations between variables above and beyond that which the unidimensional model can account for. Though only three univariate measures were investigated here, the results reported here support the conclusion that, as functions of one variable, univariate discrepancy measures are less than ideally-suited to address issues of multidimensionality.

Bivariate Discrepancy Measures

We begin with behaviors of the median PPP-values (Figures 17-24) that are common to all the discrepancy measures and then pursue differences. It is clear that moving down columns leads to more extreme median PPP-values. This supports hypothesis H1: increases in the strength of dependence lead to PPP-values that deviate from 0.5. This holds for all types of item-pairs except the pairings of items in which only one item reflects multiple dimensions (plotted in purple triangles) which do not deviate much from 0.5.

The effect of increasing the correlations among the latent dimensions is just as clear. Moving across the rows, the median PPP-values become less extreme, indicating that larger correlations lead to more moderate PPP-values. This supports hypothesis H2.

There is also evidence to support hypothesis H3. Within any panel, the medians for item-pairs that reflect the same multiple dimensions (red circles) get more moderate as the proportion of multidimensional items moves from low to medium to high. This is most evident in the lower left panel of each figure, which corresponds to the optimal combination of high strength of dependence and no correlation among the latent dimensions. Conversely, increases in the proportion of multidimensional items leads to the medians for item-pairs that reflect different multiple dimensions (green diamonds) and the medians for item-pairs that reflect the primary dimension only (black squares) becoming more extreme.

Where the discrepancy measures differ is with respect to the *direction* the median moves for the item-pairs that reflect different dimensions as the proportion of multidimensional items increases. Again, this is most easily seen in the bottom left panel of each figure. For the X^2 and G^2 discrepancy measures (Figures 17-18), the medians for the pairs of items that reflect different multiple dimensions (green diamonds) decreases as the proportion of multidimensional items increases. For the remaining discrepancy measures, the median for this type of item-pair increases as the proportion of multidimensional items increases.

The behavior of these latter discrepancy measures is consistent with conditional covariance theory. PPP-values that are larger than .5 indicate that the model is overpredicting the value of the discrepancy measure, relative to the value in the observed

data (Chen & Thissen, 1997; Habing & Roussos, 2003; Sinharay et al., in press). This is exactly what we would expect based on conditional covariance theory. In moving from a low to high proportion of multidimensional items, the projected item vectors for items that reflect different auxiliary dimensions rotate farther away from each other (compare Figures 9 and 11), leading to a decrease in the conditional association. The associations between these types of items are smaller than the model implies they ought to be; they exhibit negative local dependence.

The medians for the X^2 and G^2 discrepancy measures for item-pairs get smaller (rather than larger) because the X^2 and G^2 discrepancy measures ignore the directionality of misfit (Chen & Thissen, 1997). The covariance, log odds ratio, model-based covariance, Q_3, residual covariance, and standardized log odds ratio residual all capture the directionality. Although the X^2 and G^2 discrepancy measures detect the multidimensionality in terms of (a) item-pairs that reflect the same multiple dimensions and (b) item-pairs that reflect different multiple dimensions, they are silent as to any differences between these types of item-pairs. The remaining bivariate discrepancy measures identify the values for the item-pairs that reflect the same multiple dimensions as being underpredicted, indicating positive local dependence, and identify the values for the item-pairs that reflect different multiple dimensions as being overpredicted, indicating negative local dependence (Habing & Roussos, 2003).

Proportion of Extreme PPP-values

Figures 25-27 plot the proportions of extreme PPP-values for X^2, the logs odds ratio, and the model-based covariance. Recall these were selected to be representative of the eight bivariate discrepancy measures. The results for X^2 are representative of G^2, the

results for the model-based covariance are representative of Q_3, and the results for the log odds ratio are representative of the covariance, residual covariance, and the standardized log odds ratio residual.

Figures 25-27 reveal that, for all discrepancy measures and these types of item-pairs, power increases as the strength of dependence increases and the correlations between the dimensions decrease. This is consistent with the behavior of the medians and constitutes evidence in favor of H1 and H2.

Further, it is observed that the model-based covariance is the most powerful across all conditions, followed by the log odds ratio and then X^2. As such, hypothesis H7 was not supported. The residual covariance, log odds ratio, and the standardized log odds ratio residual were comparable to each other (and comparable to the covariance), but they were not the most effective discrepancy measures. The model-based covariance and Q_3 had the highest proportion of extreme PPP-values.

The behavior of the different types of item-pairs observed in the medians is reflected in the proportions of extreme PPP-values. Figure 25 shows that the proportion of extreme PPP-values for item-pairs that reflect the same auxiliary dimension decreases with increases in the proportion of multidimensional items. Figures 26 and 27 show that the proportions for item-pairs that reflect different auxiliary dimensions (Figure 26) and item-pairs that reflect the primary dimension only (Figure 27) increase with the proportion of multidimensional items. However, in many conditions the proportions for these latter two types of item-pairs are very low, suggesting that unless (a) the strength of dependence on the auxiliary dimensions is close to the strength of dependence on the primary dimensions and (b) the correlations among the latent dimensions are small, there

is very little power to detect the multidimensionality via these types of item-pairs. As expected, the proportions for item-pairs that reflect the same auxiliary dimension (Figure 25) are much higher, though they also decrease rapidly as the strength of dependence on the auxiliary dimension decreases and (especially) as the correlations between the latent dimensions increases.

Figures 28-30 indicate that hypothesis H4 was supported. For these types of item-pairs, increases in sample size leads to increases in the proportion of extreme PPP-values for the model-based covariance. Though not reported on space considerations, this same effect was observed for all bivariate discrepancy measures. The findings here are consistent with the pervasive finding in statistical modeling that larger samples lead to increased ability to detect model misspecifications.

Conjunctive Multidimensional Data

Median PPP-values for 2500 Examinees

Univariate Discrepancy Measures

Figures 31-33 plot the median PPP-values for the proportion correct, X^2, and G^2 discrepancy measures for the combinations of the strength of dependence, correlations among the latent variables, and (within each panel) the proportion of multidimensional items. These plots are structured like those in Figures 14-16, except the values defining the strength of dependence are those of the location parameters for the second and third dimensions in the conjunctive MIRT model rather than the values for discriminations in the compensatory MIRT model. Items reflecting the first dimension are pooled and plotted with black squares. Items reflecting both the first dimension and an auxiliary

dimension are pooled and plotted with red circles (see Appendix C). As in previous figures, lines connecting points are plotted for visual ease.

For all three discrepancy measures, across all combinations of strength of dependence, correlations among the latent dimensions, and proportion of items reflecting multiple dimensions, the red circles lie almost exactly on top of the black squares. Though there are slight deviations, no systematic patterns are present, and no deviation is large.

Figure 31: Median PPP-values for the proportion correct when the data follow a conjunctive MIRT model and N=2500.

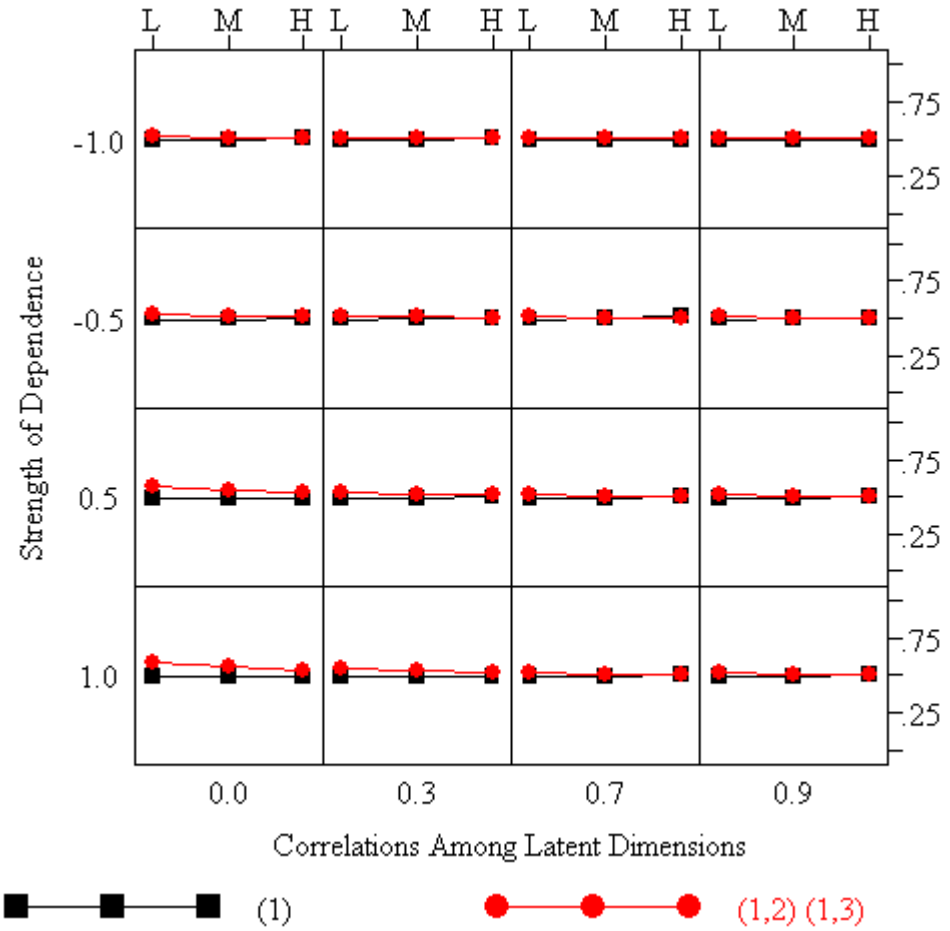


Figure 32: Median PPP-values for X^2 when the data follow a conjunctive MIRT Model and N=2500.

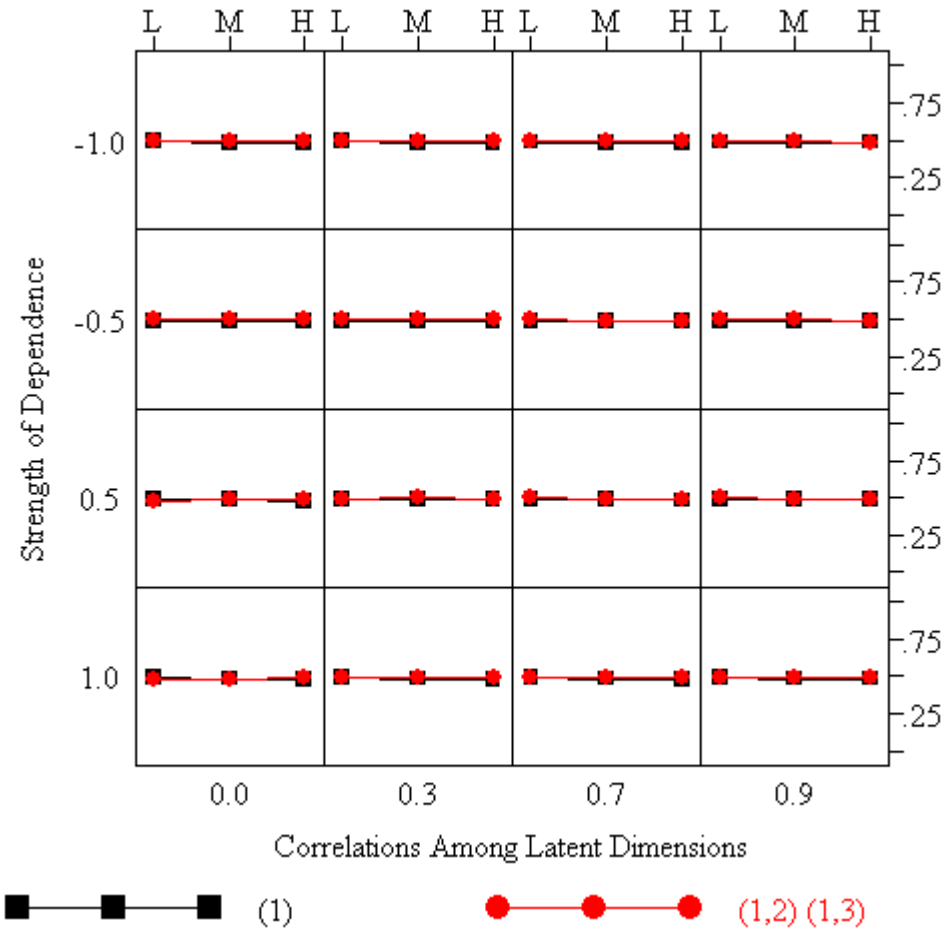
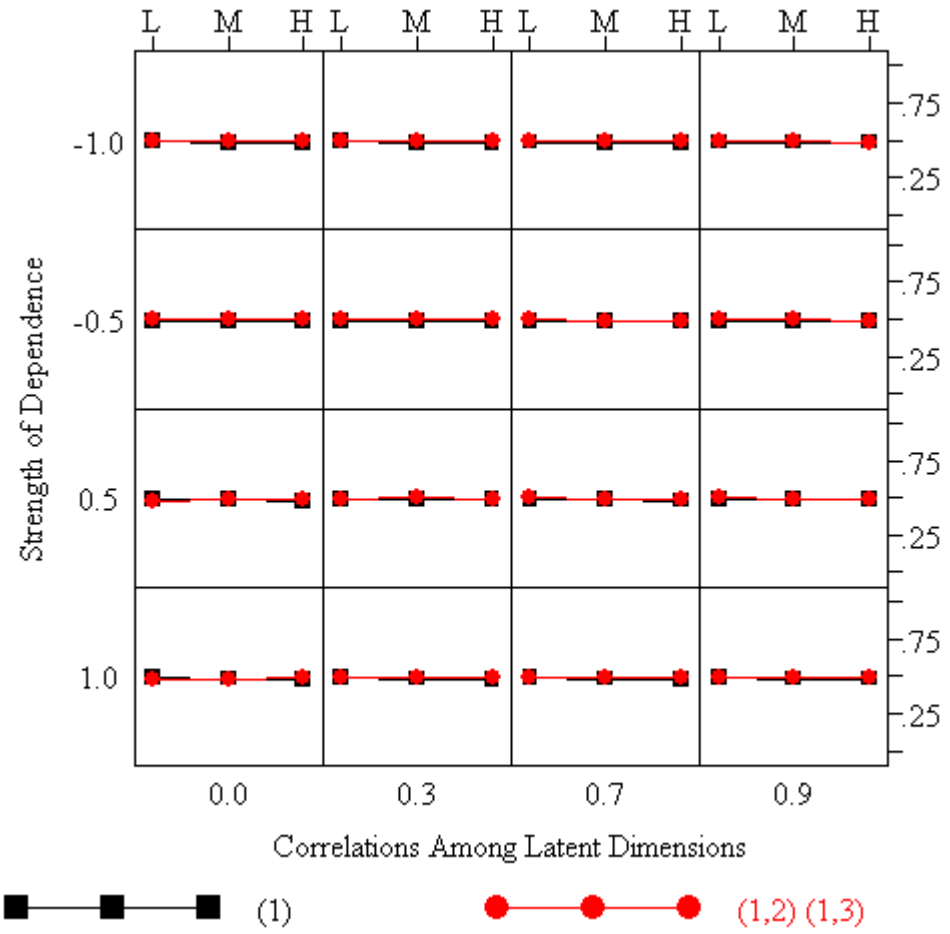


Figure 33: Median PPP-values for G^2 when the data follow a conjunctive MIRT Model and N=2500.



Bivariate Discrepancy Measures

Figures 34-41 plot the median PPP-values for the bivariate discrepancy measures for pairs of items. The structure and contents of the plots are akin to the counterparts in the analysis of compensatory multidimensional data (Figures 17-24). In all plots, item-pairs of each of the four types were pooled, reflecting exchangeability assumptions, discussed above (see also Appendix C). The results for the X^2 and G^2 discrepancy

measures (Figures 34 and 35) are quite similar. The results for the other bivariate discrepancy measures (Figures 36-41) are also similar to each other.

Figure 34: Median PPP-values for X^2 for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

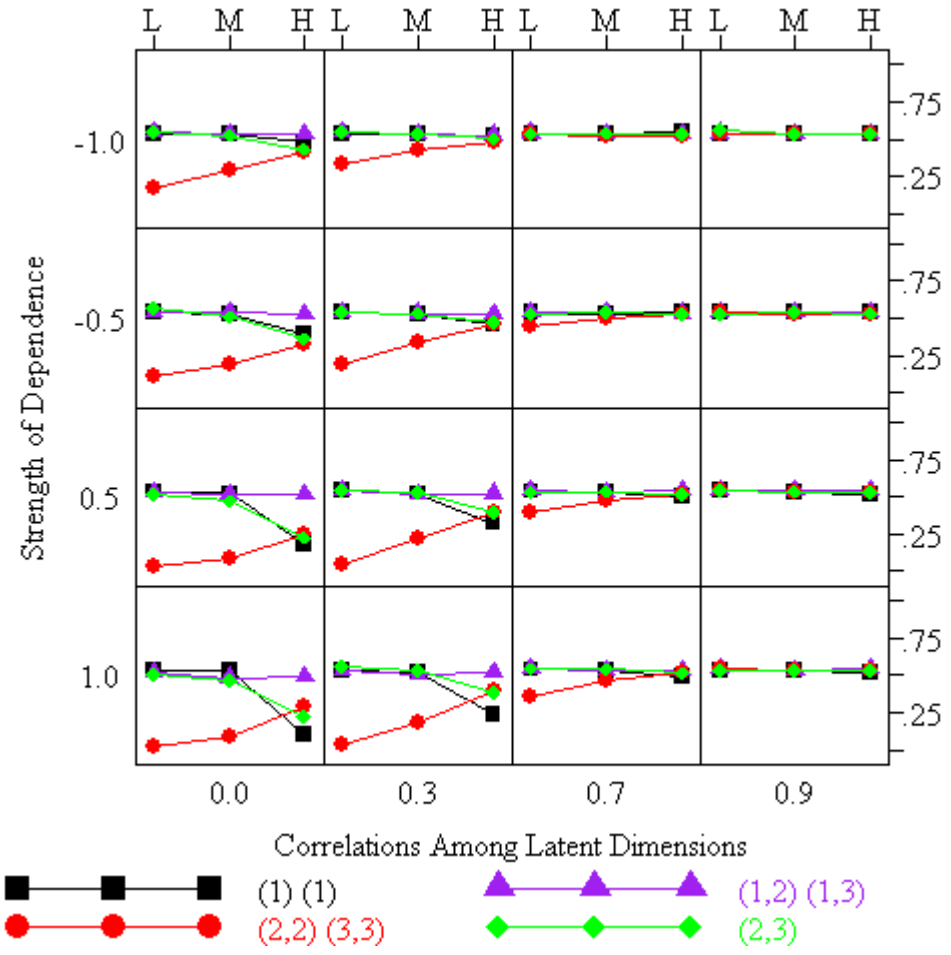


Figure 35: Median PPP-values for G^2 for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

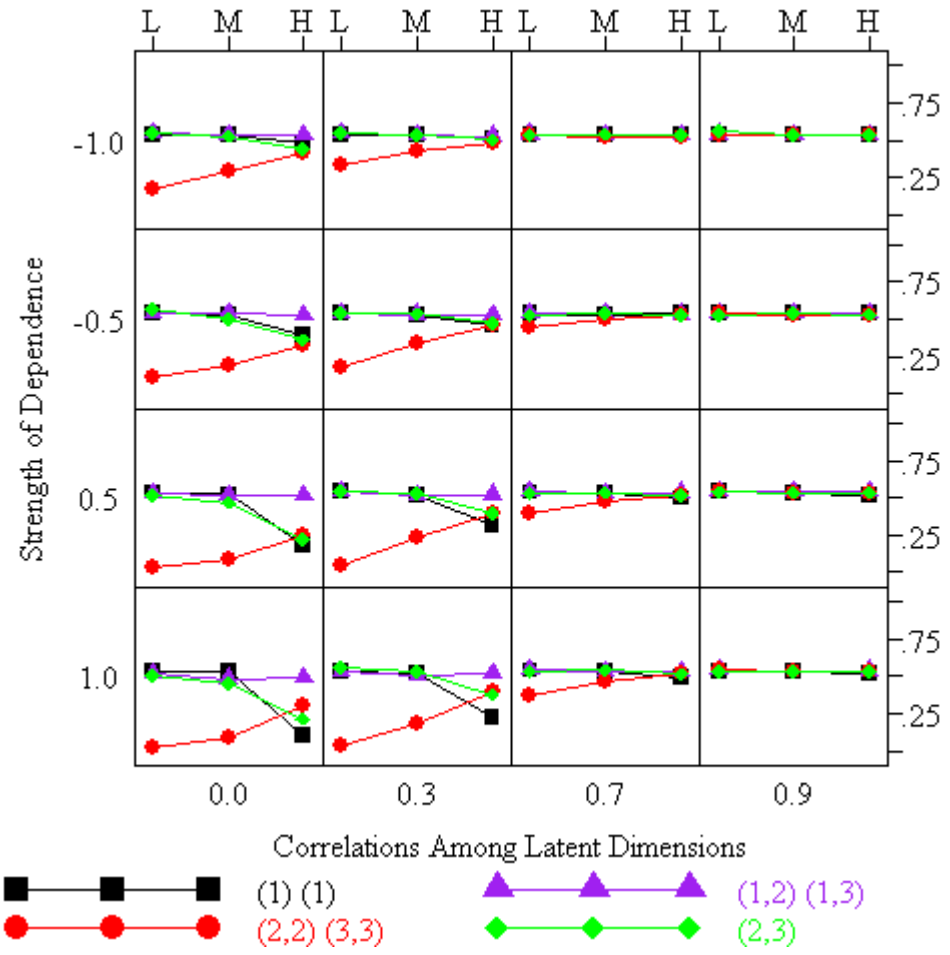


Figure 36: Median PPP-values for the covariance when the data follow a conjunctive MIRT Model and N=2500.

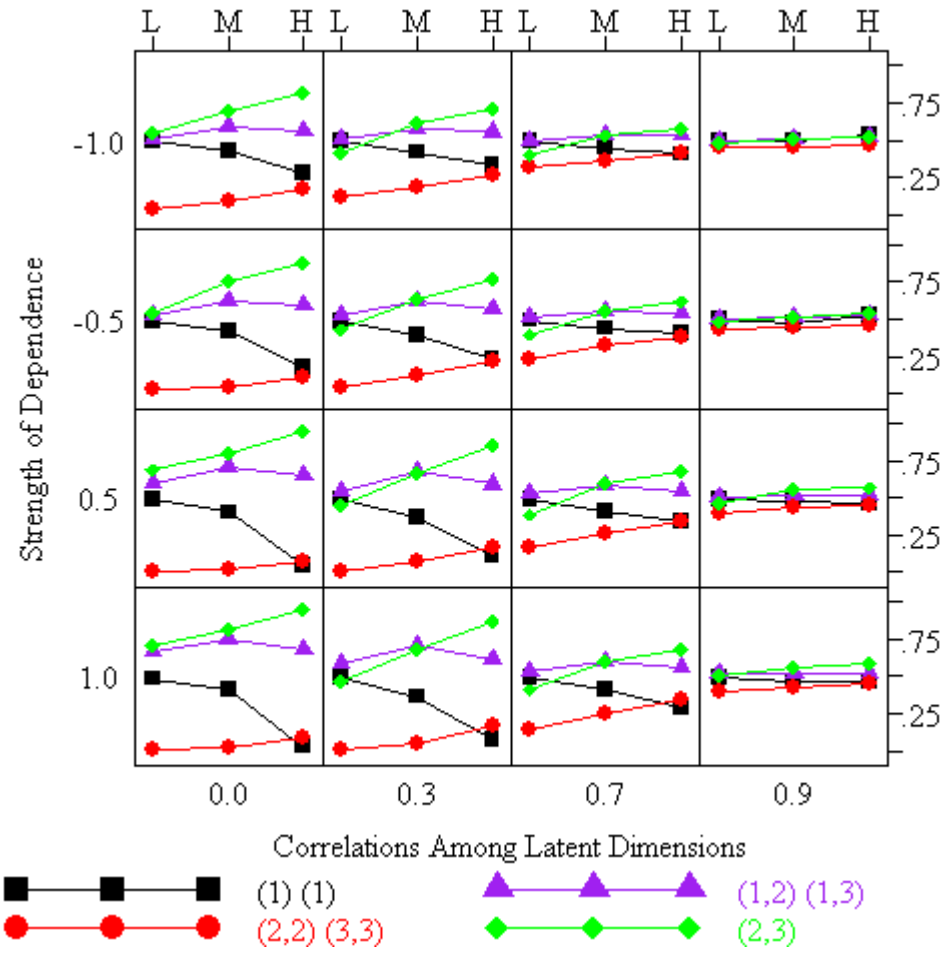


Figure 37: Median PPP-values for the log odds ratio for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

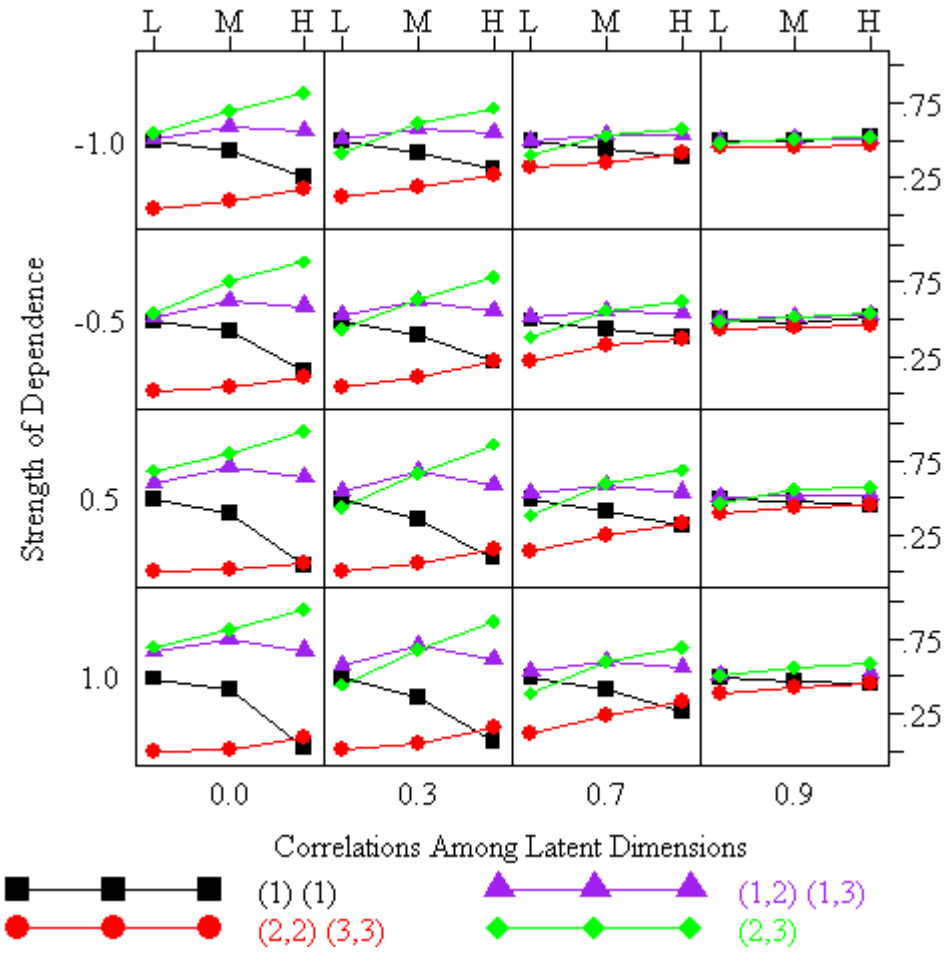


Figure 38: Median PPP-values for the model-based covariance for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

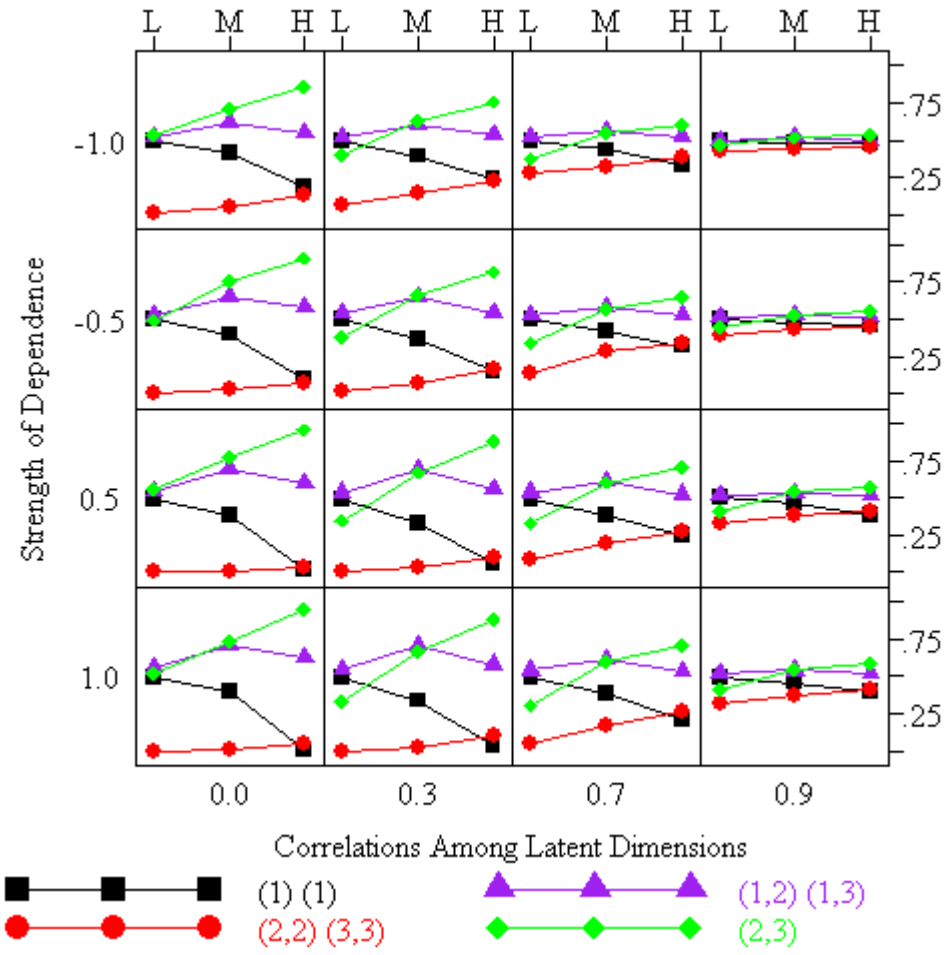


Figure 39: Median PPP-values for Q_3 for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

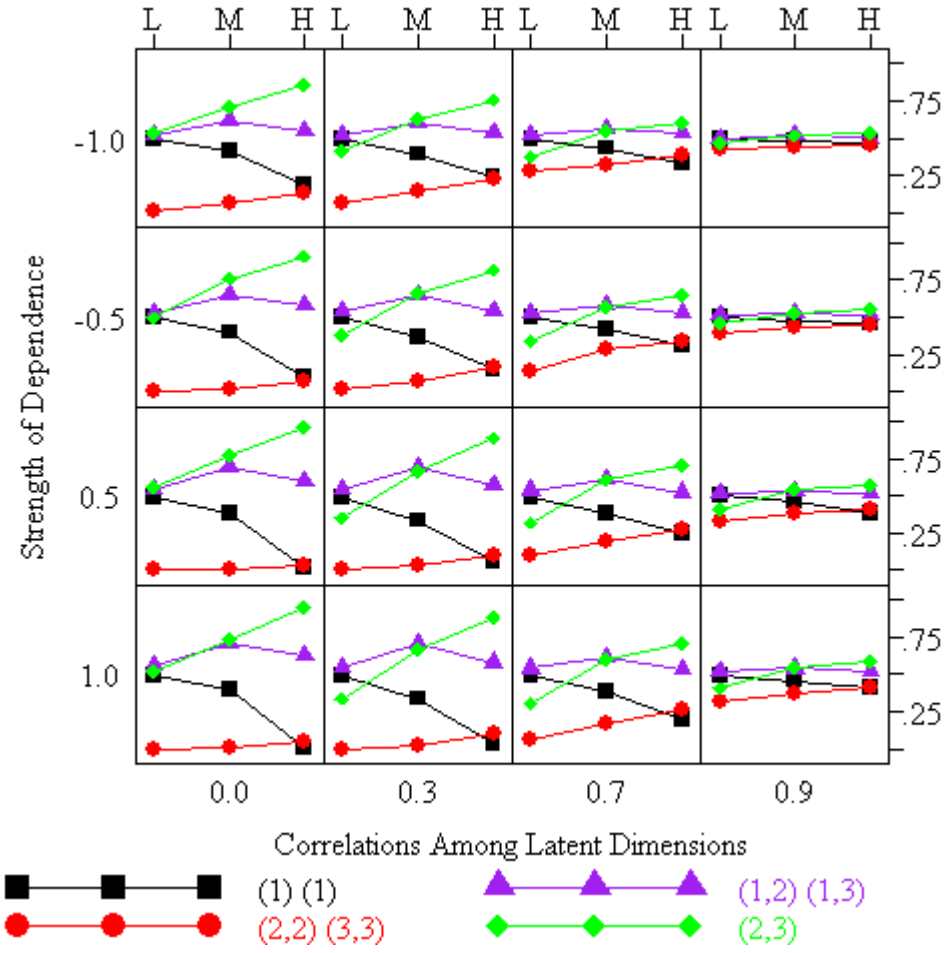


Figure 40: Median PPP-values for the residual covariance for item-pairs when the data follow a conjunctive MIRT Model and N=2500.

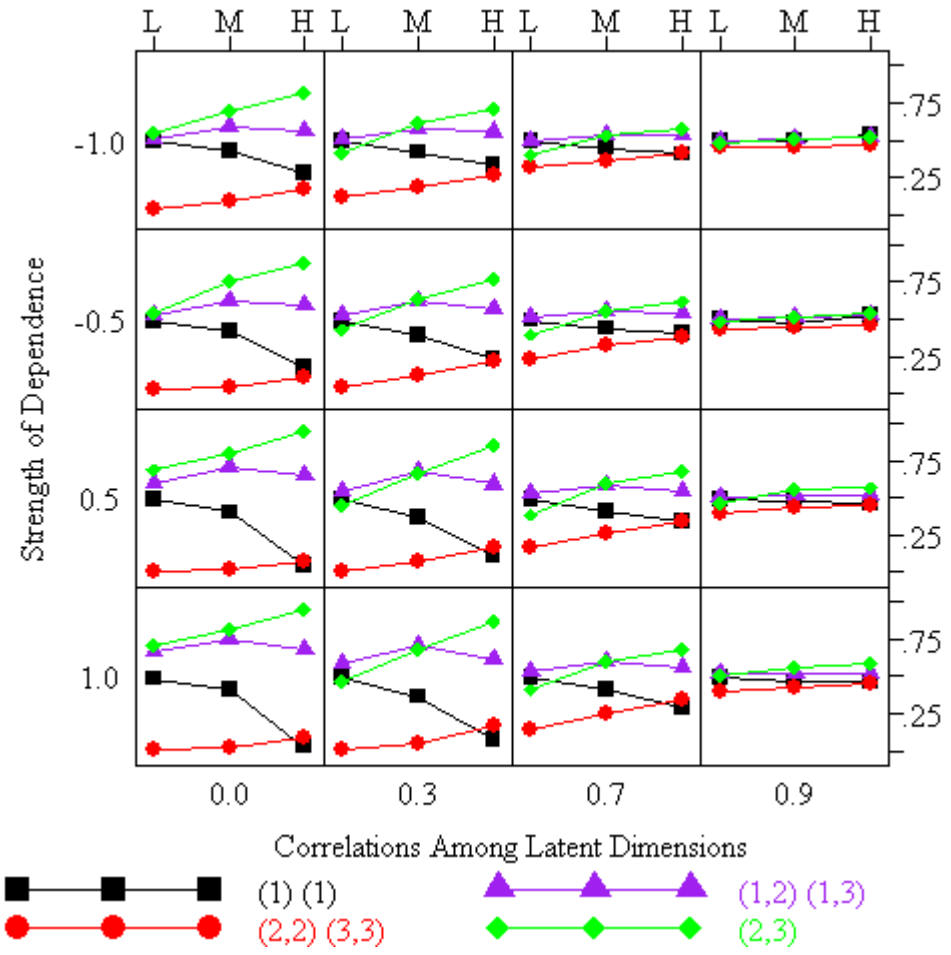
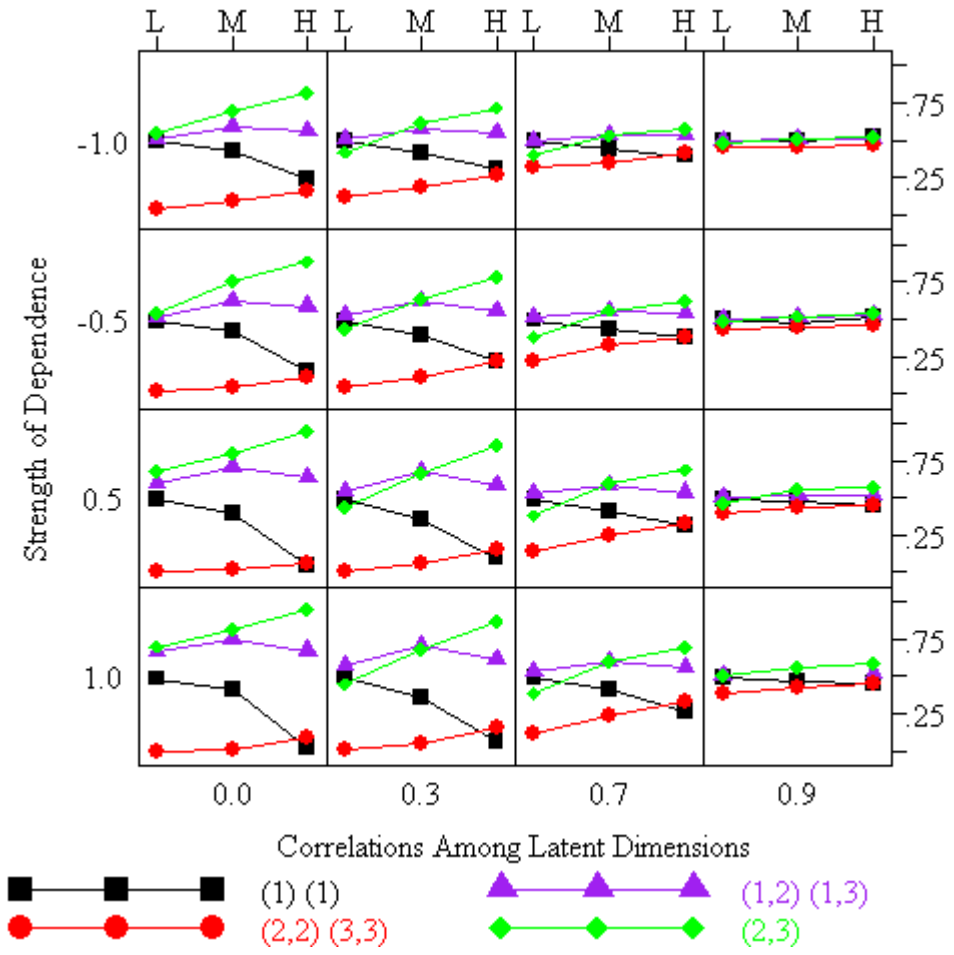


Figure 41: Median PPP-values for the standardized log odds ratio residual for item-pairs when the data follow a conjunctive MIRT Model and N=2500.



Proportion of Extreme PPP-values for 2500 Examinees

As in the previous analyses, we consider a PPP-value to be extreme when it is less than .05 or greater than .95 and confine our attention to the bivariate discrepancy measures.

Proportions of Extreme PPP-values by Type of Item-Pair

We begin with the condition in which there is a low proportion of items that reflect multiple dimensions. Table 4 presents the proportions of extreme PPP-values for

the bivariate discrepancy measures for item-pairs in which both items reflect the same multiple dimensions.

The proportions for X^2 and G^2 are quite close to one another and are always the lowest. The proportions for the model-based covariance and Q_3 are quite close to each other. One or both of these measures is always the largest, though in some conditions other measures have values equal to these. The results for the covariance and the residual covariance are close and are comparable to the results for the log odds ratio and the standardized log odds ratio residual, which themselves are quite close. These patterns of similarities held for the remaining types of item-pairs in this condition, and in the remaining levels of sample size and the proportion of items reflecting multiple dimensions. Again, graphical representations of select discrepancy measures will be presented. As was the case for the compensatory MIRT data, the results for the X^2 are representative of G^2 also; the results for the model-based covariance are representative of Q_3 also. The results for the log odds ratio are representative of the covariance, residual covariance, and the standardized log odds ratio residual.

Table 4: Proportion of replications with extreme PPP-values (i.e., PPP-value < .05 or >.95) for item-pairs that reflect the same multiple dimensions when data follow a conjunctive MIRT model, N=2500, and the proportion of items is low.

		Discrepancy Measure							
b_{j2}, b_{j3}	ρ	X^2 Pair	G^2 Pair	Cov	LN (OR)	Mod-Based Cov	Q_3	Resid Cov	Std LN(OR) Resid
-1.0	0.0	.23	.23	.58	.57	.70	.70	.58	.59
	0.3	.06	.06	.30	.29	.45	.45	.30	.29
	0.7	.02	.02	.09	.09	.13	.12	.09	.09
	0.9	.01	.01	.03	.03	.07	.07	.03	.03
-0.5	0.0	.36	.35	.65	.64	.74	.74	.65	.64
	0.3	.23	.24	.51	.51	.69	.69	.51	.51
	0.7	.04	.04	.17	.18	.28	.27	.17	.18
	0.9	.01	.01	.04	.04	.13	.13	.04	.04
0.5	0.0	.60	.59	.84	.86	.93	.93	.84	.87
	0.3	.52	.51	.75	.78	.88	.88	.75	.76
	0.7	.02	.02	.15	.18	.34	.36	.15	.18
	0.9	.00	.00	.00	.02	.08	.08	.00	.02
1.0	0.0	.65	.62	.83	.84	.93	.93	.83	.84
	0.3	.56	.55	.74	.80	.91	.92	.74	.80
	0.7	.02	.02	.20	.24	.50	.49	.20	.23
	0.9	.00	.00	.02	.03	.12	.12	.02	.04

The panels in Figure 42 plot the proportions of extreme PPP-values for item-pairs that reflect the same auxiliary dimension. Figures 43 and 44 plot the proportions for item-pairs that reflect different auxiliary dimensions and item-pairs that 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 did not meaningfully deviate from .5 for any of the discrepancy measures.

Figure 42: 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 the data follow a conjunctive MIRT Model and N=2500.

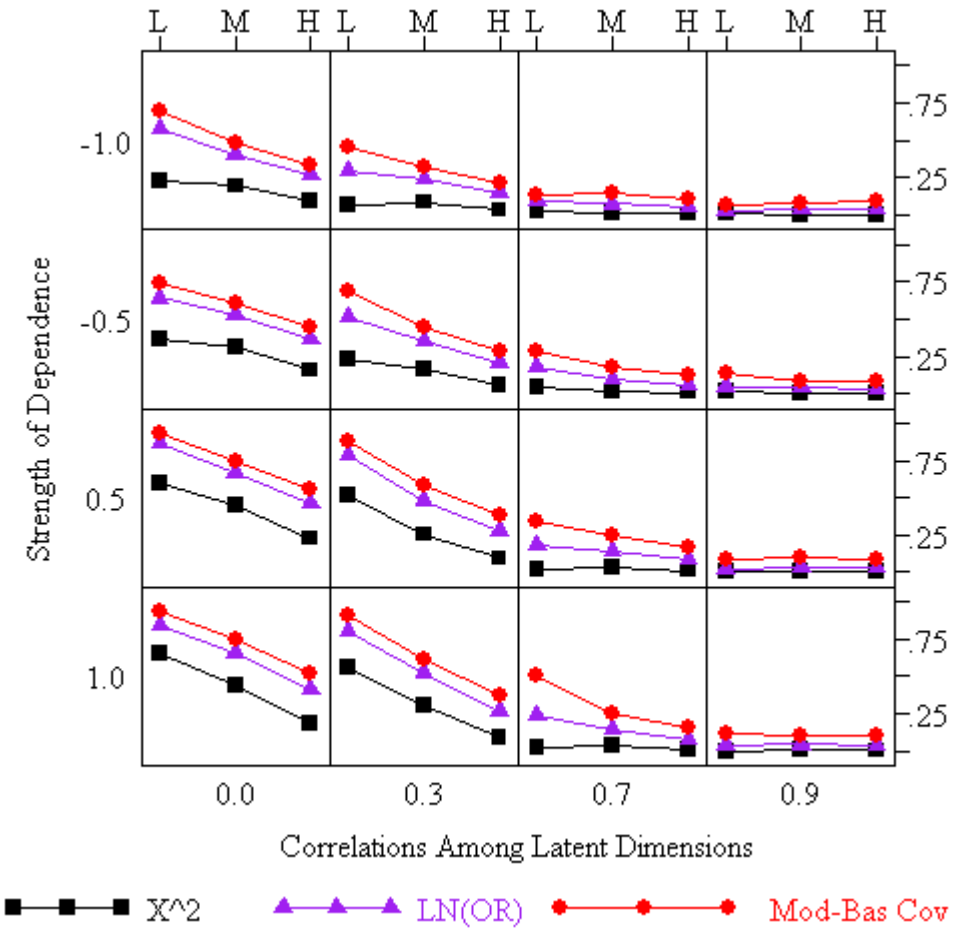


Figure 43: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for select discrepancy measures for item-pairs that reflect different multiple dimensions when the data follow a conjunctive MIRT Model and N=2500.

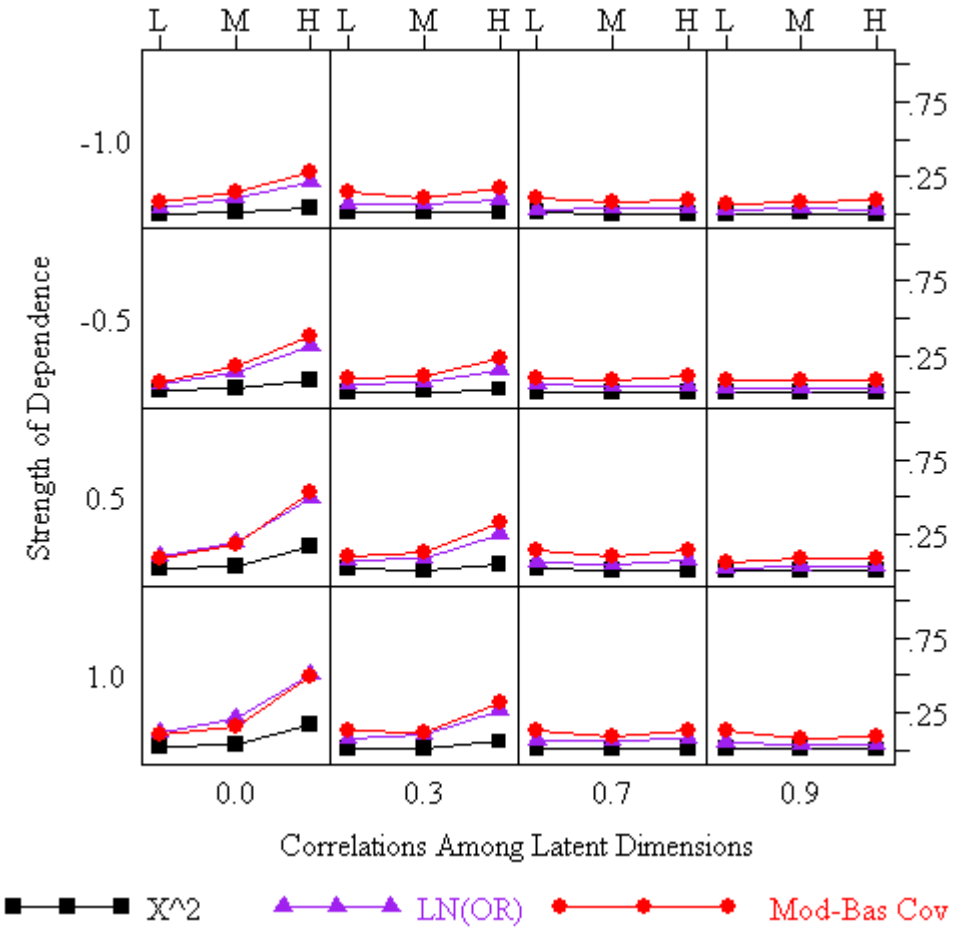
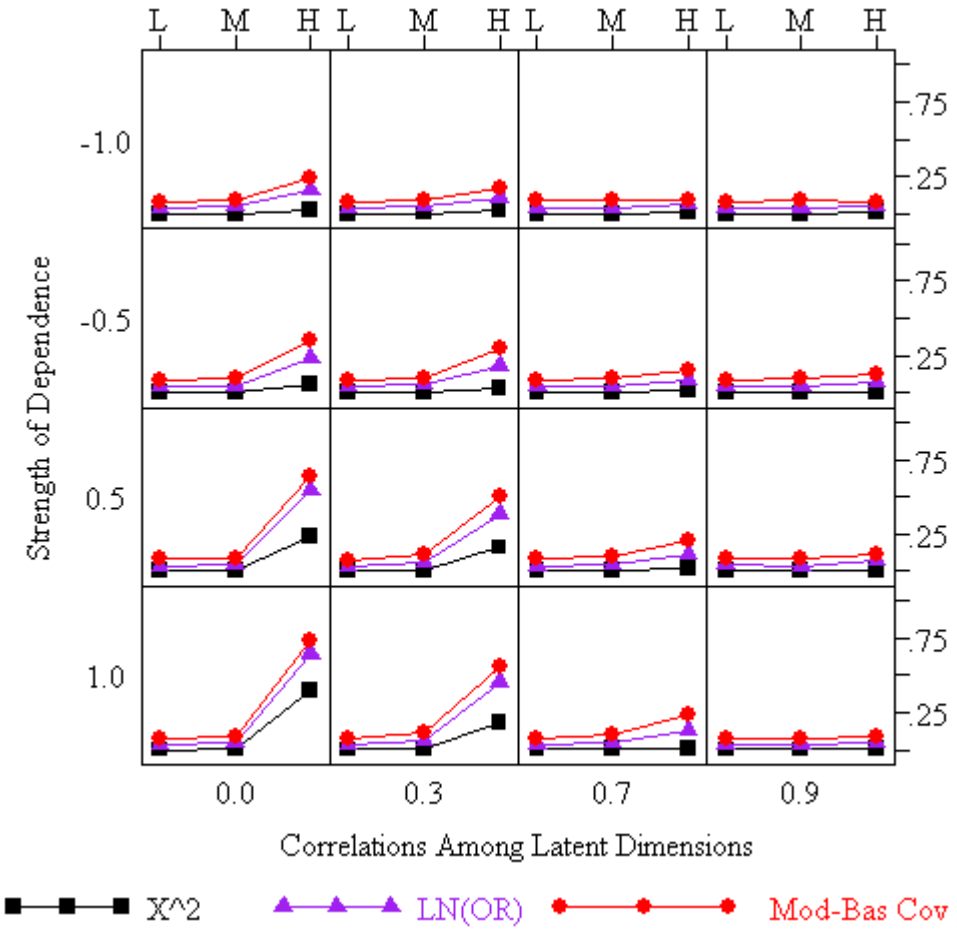


Figure 44: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for select discrepancy measures for item-pairs that reflect the primary dimension only when the data follow a conjunctive MIRT Model and N=2500.



Proportions of Extreme PPP-values by Sample Size

As in the case of compensatory data, we employ the model-based covariance to illustrate the influence of sample size. Figures 45-47 display plots of proportion of extreme PPP-values for item-pairs that reflect (a) the same multiple dimensions, (b) different multiple dimensions, and (c) the primary dimension only, respectively, at each condition.

Figure 45: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect the same multiple dimensions when the data follow a conjunctive MIRT Model.

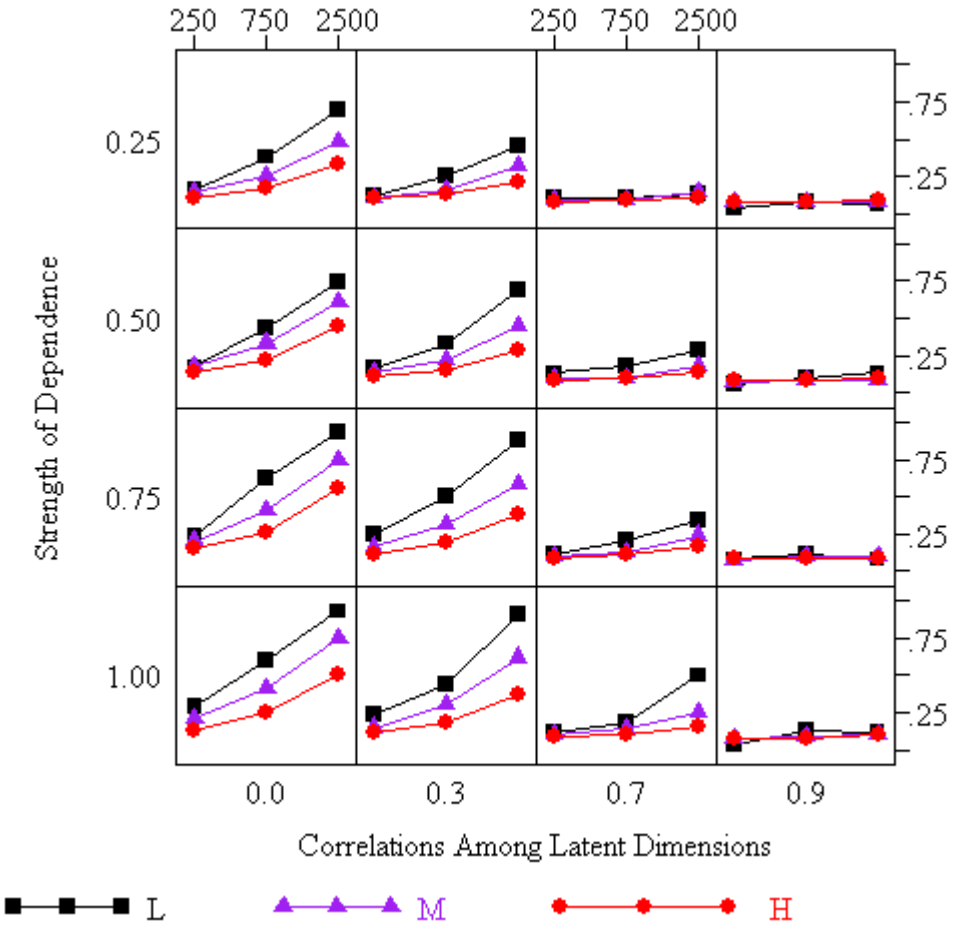


Figure 46: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect different multiple dimensions when the data follow a conjunctive MIRT Model.

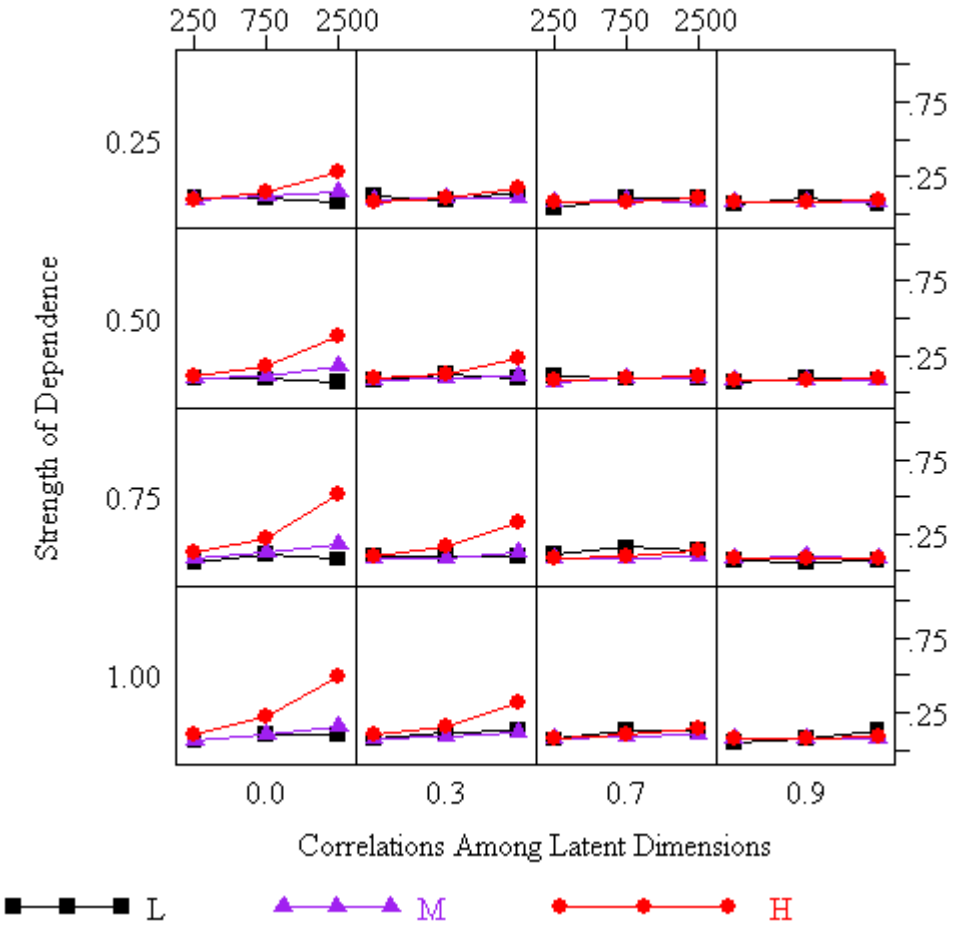
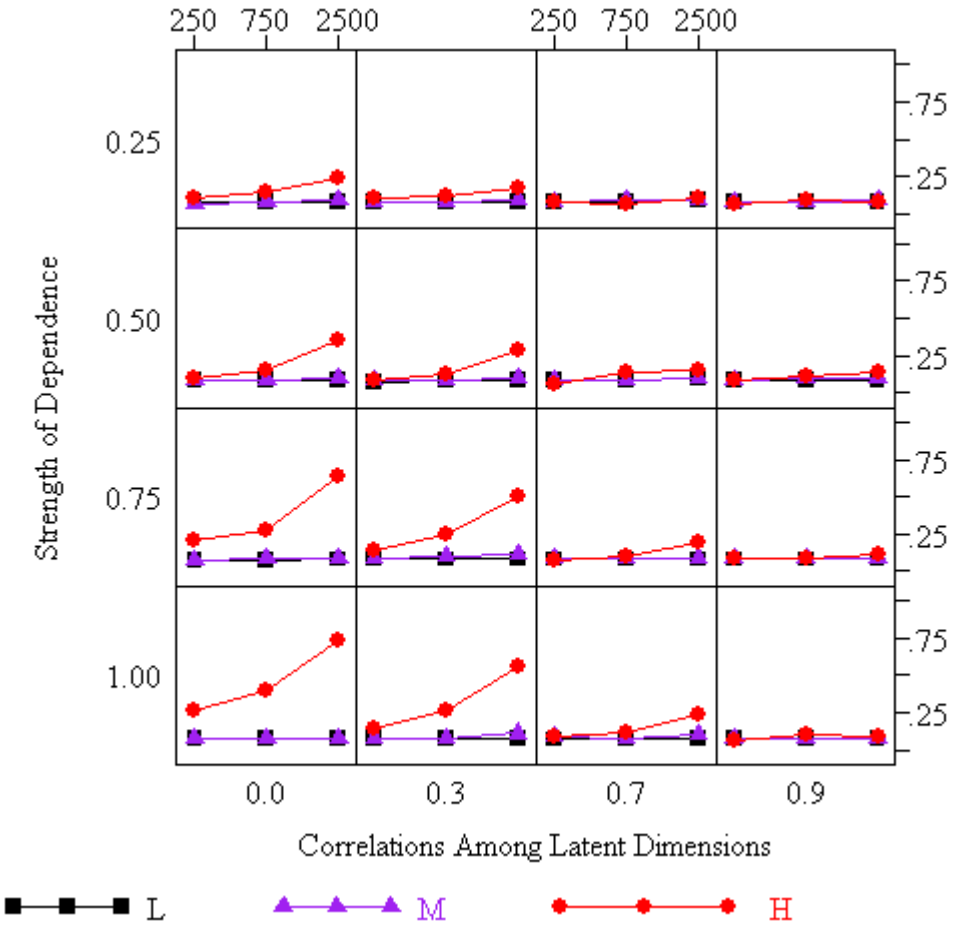


Figure 47: Proportion of extreme PPP-values (i.e., PPP-value < .05 or >.95) for the model-based covariance for item-pairs that reflect the primary dimension only when the data follow a conjunctive MIRT Model.



Discussion

Univariate Discrepancy Measures

Beginning with the univariate discrepancy measures (Figures 31-33), it is observed that for all levels of the strength of dependence and all levels of the correlations between the dimensions, the circles are essentially at the same points as the squares. Substantively, the proportion correct, X^2, and G^2 discrepancy measures were unable to

detect the multidimensionality in any of the conditions. This is consistent with the analysis of compensatory data and supports hypotheses H5 and H6.

Bivariate Discrepancy Measures

Turning to the bivariate discrepancy measures, we discuss general patterns exhibited across the discrepancy measures before explicating their differences. Consistent with the findings based on the compensatory data, consideration of Figures 34-41 supports hypotheses H1 and H2. Increases in the strength of dependence on an auxiliary dimension leads to PPP-values farther from .5. Increases in the correlations among the dimensions leads to PPP-values closer to .5. This holds for all types of item-pairs except the pairings of items that reflect different multiple dimensions (plotted in purple triangles) which do not deviate much from 0.5.

Hypothesis H3 is also supported. Within any panel, the medians for item-pairs that reflect the same multiple dimensions (red circles) approach .5 as the proportion of multidimensional items moves from low to medium to high. Conversely, increases in the proportion of multidimensional items leads to the medians for items pairs that reflect different multiple dimensions (green diamonds) and the medians for item-pairs that reflect the primary dimension only (black squares) deviating farther from .5.

As was found in the analysis of compensatory data, the X^2 and G^2 discrepancy measures (Figures 34 and 35) are insensitive to the direction of misfit (Chen & Thissen, 1997) and therefore fail to identify the negative local dependence between items that reflect different auxiliary dimensions. The remaining discrepancy measures (Figures 36-41) all reflect this directionality and do not differ greatly from one another in terms of the median PPP-values across the conditions.

Proportion of Extreme PPP-values

Figures 42-44 plot the proportions of extreme PPP-values for X^2, the logs odds ratio, and the model-based covariance, which are representative of the eight bivariate discrepancy measures. For all discrepancy measures and types of item-pairs, the proportion of extreme PPP-values increases with the strength of dependence and decreases with the correlations among the latent dimensions, supporting hypotheses H1 and H2.

Several other patterns consistent with the compensatory data emerge. In contrast to hypothesis H7, the model-based covariance, which is also representative of Q_3, exhibited the highest proportion of extreme PPP-values across almost all conditions and all types of item-pairs. Exceptions to this were rare and trivial. For example consider the pairings of items that reflect different multiple dimensions when the strength of dependence is highest, the latent dimensions are uncorrelated, and the proportion of multidimensional items is medium (i.e., the middle point in the lower left panel of Figure 43). It can be seen that the model-based covariance (and Q_3,) lags slightly behind that of the log odds ratio (i.e., the red circle is just slightly lower than the purple triangle).

Also, the proportions of extreme PPP-values for item-pairs that reflect the same auxiliary dimension (Figure 42) decrease as the proportion of multidimensional items increases. The proportions for the other types of item-pairs (Figures 43-44) increase as the proportion of multidimensional items increases.

For item-pairs in which both items reflect the primary dimension only (Figure 44) under certain combinations of the strength of dependence and the correlations among the dimensions, there is a spike when the proportion of multidimensional items is high. In

the current case, PPMC fails to consistently pick up on the multidimensionality for low and medium proportions of multidimensional items, but improves dramatically when the proportion of multidimensional items is large. A milder effect of this sort was observed in the analysis of the compensatory MIRT data.

Figure 45, and to a lesser extent, Figures 46-47 indicate that, as expected, increase in sample size leads to an increase in the proportion of extreme PPP-values. For item-pairs that reflect the same multiple dimensions (Figure 45), increasing sample size has the usual effect of increasing the power to detect data-model misfit across all conditions except those in which the correlations between the dimensions are .9 and the condition in which the correlations between the dimensions are .7 and the strength of dependence is weakest.

For item-pairs that reflect different multiple dimensions (Figure 46) and for item-pairs that only reflect the primary dimension (Figure 47), sample size only exerts an influence when the proportion of multidimensional items is high. What's more, the proportions of extreme PPP-values are larger for the item-pairs that reflect the primary dimension only than for item-pairs that reflect different multiple dimensions (i.e., the red circles in the left two columns in Figure 47 are higher than their counterparts in Figure 46). This stands in sharp contrast to the results for the compensatory data, in which the items that reflected different multiple dimensions had a higher proportion of extreme PPP-values (Figures 29 and 30).

Supplemental Analyses

In exploring the tenability of the exchangeability assumptions with respect to pooling the results from both auxiliary dimensions (Appendix C), an unanticipated

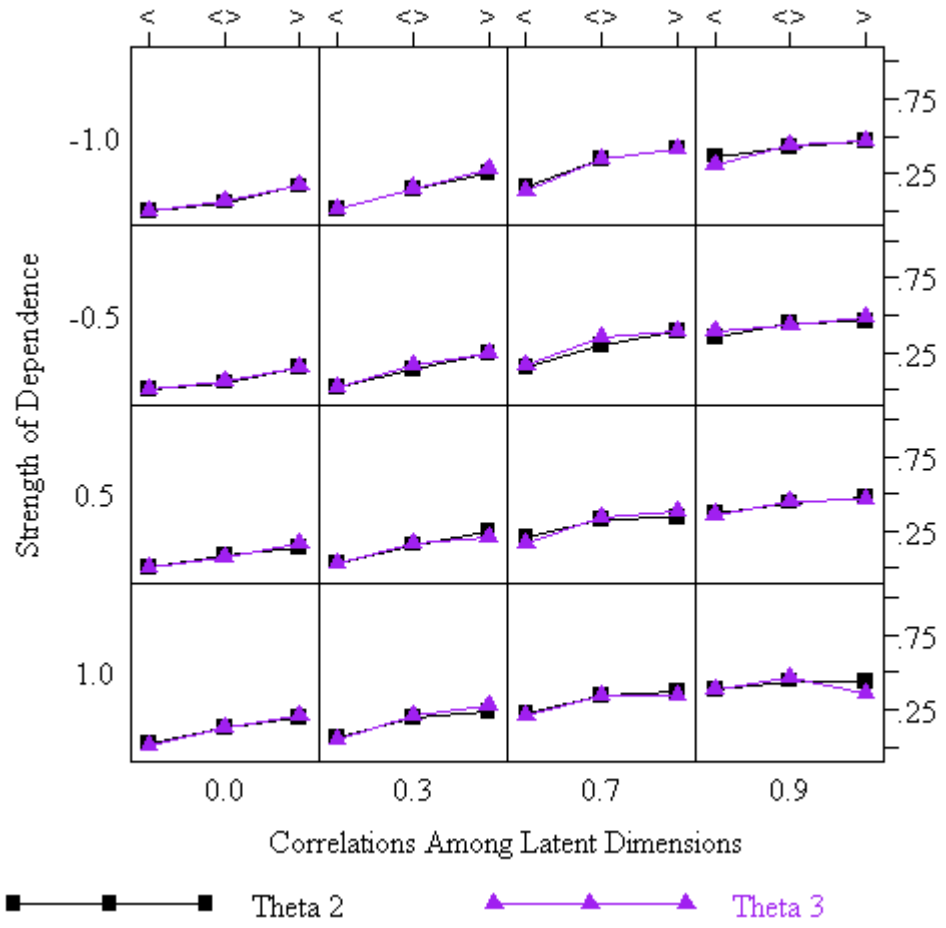
relationship between the item difficulty along the primary dimension and the efficacy of PPMC emerged. The exchangeability analyses suggest that, in addition to the factors studied, the performance of PPMC also depended on the location of the items along the *primary* dimension (Appendix C). Pooling all item-pairs of the same type (e.g., all item-pairs that reflect the first and second dimension) might be unwarranted. Rather, item-pairs may need to be separated in terms of the difficulty along the first dimension.

Additional analyses pursuing these effects were conducted. They will not be presented in full here on considerations of space. As an example, Figure 48 plots the median PPP-values for the model-based covariance based on conjunctive MIRT when the proportion of multidimensional items is high and sample size is 2500. The rows and columns of the plot define the combinations of the strength of dependence and the correlations among the dimensions.

Within each panel, the three points on the horizontal axis characterize the pairing of items in terms of their locations along the primary dimension. The left point, labeled as '<', refers to pairs of items in which both item difficulties along the primary dimension are less than the item difficulty along the second or third dimension. The middle point, labeled as '< >', refers to pairs of items in which one item has a difficulty parameter along the first dimension below the difficulty along the second or third dimension and the other item has a difficulty parameter along the first dimension above the difficulty along the second or third dimension (i.e., in terms of the difficulties along the first dimension, the items "straddle" the difficulty along the auxiliary dimension). The right point, labeled as '>', refers to pairs of items in which both items have difficulty parameters

along the first dimension that are above the difficulty along the second or third dimension.

Figure 48: Median PPP-values for the model-based covariance for item-pairs that reflect the same multiple dimensions for conjunctive MIRT data.



In each panel, the black squares are the median PPP-values for item-pairs that reflect θ_1 and θ_2 and the purple triangles are the median PPP-values for item-pairs that reflect θ_1 and θ_3. Note that there are only trivial differences between the item-pairs that reflect the different auxiliary dimensions. However, there is a noticeable trend for the sets of points in which the median PPP-values are more moderate for the pairings of

items with difficulty parameters along the primary dimension above the difficulty parameters along the auxiliary dimensions.

A possible explanation for this effect comes from considering what occurs when we move from a unidimensional model to a conjunctive MIRT model. We can think of such a transition as the inclusion of the requirement of certain levels of proficiency of auxiliary dimensions. In other words, a unidimensional model states that a certain level of proficiency is needed to have high probability of solving the item. A conjunctive model states that, *in addition*, a certain level of proficiency is needed along another dimension in order to *retain* the high probability of solving the item. For items that are *difficult* in terms of the primary dimension (for a given population of examinees), the inclusion of a conjunctive effect with an auxiliary dimension has a mild impact. These items are difficult to begin with. For examinees with proficiencies below the difficulty on the primary dimension, requiring proficiency along another dimension has little influence, as they had smaller probabilities of solving the item anyway.

For items that are *easier* in terms of the primary dimension, the inclusion of a conjunctive effect is more influential on the probabilities of correctly solving the items. If only the primary dimension were relevant, many examinees would be solving such items. The incorporation of another dimension effectively requires much more of the examinee when the difficulty parameter along this auxiliary dimension is high. In other words, items that are easy along the primary dimension (relative to their difficulty along the auxiliary dimension) are no longer solved correctly as often. The model checking is more sensitive to these items, as they are the ones more drastically altered by the inclusion of auxiliary dimensions with conjunctive effects. This explanation is

speculative; further research that explicitly studies the (potential) relationship between difficulty along the primary dimension relative to difficulty along auxiliary dimensions is needed.

SYNTHESIS AND CONCLUSIONS

Some general conclusions can be made regarding the research hypotheses and the efficacy of conducting PPMC for investigating multidimensionality in IRT.

Conditional Covariance Theory

Conditional covariance theory was developed in the context of generalized compensatory multidimensionality (Zhang & Stout, 1999a). Of considerable interest is that its implications also were observed in the analysis of conjunctive MIRT data. This suggests that the principles of conditional covariance theory may hold in situations not covered by the formal theory and that it provides a framework for treating a broader class of multidimensional models than has so far been established. Further theoretical and empirical work is needed to pursue this possibility.

PPMC for Dimensionality Assessment

The univariate discrepancy measures were found to be wholly *ineffective* for detecting the multidimensionality, across any of the conditions, supporting hypotheses H5 and H6. These findings are consistent with findings of Fu et al. (2005). Ignoring the multidimensionality amounts to ignoring sources of association. Hence, multivariate measures that reflect the degree to which the model accounts for the associations among variables are more successful than univariate measures. The proportion correct measure

also suffers in that the inclusion of a unique location parameter in the 2-PL ought to be sufficient to recover the proportion correct regardless of the misspecification of the model. Though not the focus of this study, it is maintained as tenable that the proportion correct is not useful for criticizing models with unique location parameters (Sinharay et al., in press).

The X^2 and G^2 measures for item-pairs behaved similarly to one another, and were less effective than the other bivariate measures. In addition, these measures are non-directional (Chen & Thissen, 1997), and as a consequence failed to distinguish between item-pairs that exhibited positive local dependence from those that exhibited negative local dependence (Habing & Roussos, 2003). As illustrated in the second study, differentiating between item-pairs that exhibit positive and negative local dependence may prove useful in substantive criticism and model reformulation in terms of inferring which items reflect different, unmodeled dimensions.

For the remainder of the discrepancy measures, hypotheses H1, H2, H3, and H4 were supported. PPMC improves as (a) the strength of dependence on auxiliary dimensions increases, (b) the correlations between the latent dimensions decrease, (c) the proportion of multidimensional items decreases (for items that reflect the same multiple dimensions), and (d) sample size increases. The effects for the strength of dependence, correlations among the latent dimensions, and sample size 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. Also, when the proportion of multidimensional items increases, though it becomes more difficult to detect the multidimensionality in terms of item-pairs that reflect the same auxiliary dimension, it

becomes easier to detect misspecifications in terms of item-pairs that reflect different auxiliary dimensions or item-pairs that reflect the primary dimension only, supporting hypothesis H3.

These main effects were not present in all combinations of the remaining conditions. In some cases, the effects of one condition served to mitigate the effects of others. For example, once the correlations between the dimensions got extremely strong (.9), the remaining factors became almost irrelevant. Even at the strongest levels of dependence and the largest sample size, it became virtually impossible to detect the multidimensionality.

A pervasive finding is the effect of the correlations among the dimensions. For example, examination of Figures 27-30 suggests that the influence of the correlations among the dimensions on PPMC is greatest in between 0.3 and 0.7. In other words, when the correlations among the latent dimensions are relatively extreme (large or small) in magnitude, PPMC becomes (relatively) stable in the sense that the correlations becoming *more* extreme do not seem to have much more of an effect. However, if the correlations among the dimensions are more moderate in magnitude, slight changes in the correlations have more of an impact on PPMC. Further research that specifically targets these values of the correlations is needed.

Turning to the discrepancy measures themselves, the covariance, log odds ratio, model-based covariance, Q_3, residual covariance, and the standardized log odds ratio residual behaved similarly to one another across conditions in terms of the *patterns* of the PPP-values and their relationships to the manipulated factors. Differences between some of these measures were observed in terms of the *magnitudes* of the PPP-values. More

specifically, 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's 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 the model-based covariance and Q_3, which performed almost identically. Their superiority is at first evident in the analysis of unidimensional data in which their distributions of PPP-values are closest to uniform. These findings suggest that the use of these discrepancy measures to conduct hypothesis testing might not lead to overly conservative inferences.

In terms of detecting multidimensionality, these measures outperformed the others in all the conditions and types of item-pairs in the compensatory and almost all of the conditions and types of item-pairs in the conjunctive data; in the rare cases where they were not the best, they were not far behind. Thus hypothesis H7 was not supported. In this study the model-based covariance and Q_3 performed as well or better than the other measures. The model-based covariance and Q_3 are therefore the recommended measures for conducting PPMC to investigate unaccounted for multidimensionality. There is no empirical basis in these results for promoting one of these above the other; they often performed identically. When there were differences, neither one consistently outperformed the other and the differences were trivial.

The success of the model-based covariance and Q_3 is a somewhat surprising result. As linearly-based measures of association, the expectation was that these measures would *not* perform as well as those derived from measures of association for dichotomous variables. One possible explanation for their success concerns the way the

expected value terms in these measures were derived. Problems with linear measures of association for dichotomous variables may in part be traced to the distribution of deviation scores. For a dichotomous variable with a fixed mean, there are only two deviation scores. In contrast, a continuous variable has (in theory) an infinite number of possible deviation scores.

Returning to the model-based covariance and Q_3 studied here, note that the expectation component in each deviation score in Equations (14) and (15) is indexed by subjects (i). There is an examinee-specific expectation for each item, derived from the IRF. For a given item, these expectations may vary over a continuum because the values of the examinees' latent variables are modeled as continuous variables. Despite the fact that there remain only two possible values of the observable, there are an infinite number of expectations and those expectations will vary from examinee to examinee. The result is that there are an infinite number of deviation scores. From this perspective, even though the variable is dichotomous, calculating the model-based covariance and Q_3 is akin to calculating a covariance-based measure for continuous data.

PPMC More Generally

Consistent with theoretical and empirical findings (Robins et al., 2000; Sinharay et al., in press) all discrepancy measures showed empirical proportions of extreme PPP-values below nominal levels under null conditions. The model-based covariance and Q_3 performed best in this context, coming quite close to nominal levels. These findings suggest that the use of these discrepancy measures might not lead to overly conservative inferences.

As a caveat, it should be noted that although the hypotheses and conclusions for the discrepancy measures were formulated separately from those for the manipulated factors, the two aspects of this research are intimately linked. For example, the assertion that it becomes easier to detect multidimensionality as the strength of dependence increases is dependent on the discrepancy measure chosen. To see this, consider what the conclusions would be if the proportion correct was the only discrepancy measure investigated. Strictly on the basis of the results for the proportion correct, it might have been natural to infer that strength of dependence has no effect on our ability to detect multidimensionality. Extending this further, had the proportion correct been the only discrepancy measure, we might have concluded that *none* of the manipulated factors influenced our ability to detect multidimensionality. We are able to see the effects of the manipulated factors because of the usefulness of certain discrepancy measures.

Likewise, the assertion that the model-based covariance and Q_3 are the best discrepancy measures is linked to the manipulated factors. Had we only investigated conditions in which the correlations among the dimensions were extremely positive, we would not be justified in declaring certain measures as preferred over others.

This connection between the manipulated factors and the discrepancy measures underscores a more general point about PPMC. Careful thought must be given to the choice of discrepancy measures. Choices ought to be guided by *statistical and substantive* considerations. As discussed above, consideration of the proportion correct and the 2-PL suggests that the proportion correct will not be successful. Consideration of the hypothesis of multidimensionality suggests that bivariate discrepancy measures aimed at capturing how well the model accounts for associations among the variables ought to

outperform univariate discrepancy measures and bivariate measures that ignore direction. The appropriate choice of discrepancy measures is therefore contingent on both the model and hypotheses regarding potential data-model misfit.

CHAPTER 5: POSTERIOR PREDICTIVE MODEL CHECKING FOR MULTIDIMENSIONALITY IN BNS

The first study investigated PPMC for criticism of unidimensional IRT models. Familiar IRT models may be taken "off the shelf" and used in operational assessment when theory and data structures align with the assumptions of the model. In nonstandard environments, the straightforward application of these models will not suffice. This is particularly true when the domain consists of multiple relevant variables, some or all of which may be targets of inference, or when evidence comes in nonstandard forms or is related to the desired inference(s) in atypical ways.

Rather, what are needed are flexible model-construction strategies that may be re-instantiated in different settings to build models tuned to the salient features of the desired inference(s) and potential evidence (Rupp, 2002). To this end, a more general model-building perspective seeks to construe theory, models, and data jointly (Mislevy, Steinberg, & Almond, 2003). Models are viewed as tools for structuring the multiple, possibly complex relationships among variables characterizing the target inferences and potential evidence (Pearl, 1988). In terms of the focus of this work, assumptions from theory that may be expressed as exchangeability structures (de Finetti, 1964) imply conditional independence relationships alleged to hold in the data (see Pearl, 1988, for further details on connections between statistical matters such as conditional independence and the psychology of human inference).

In light of framing the theory, model, and data jointly, results may lead to revision of any of the three. For example, unexpected relationships in the data may necessitate changes in the model and or suggest alternative theoretical explanations. The point is

that rather than *choose* a model from an existing set, we should *build* one in conjunction with the development of the substantive theory and intended interpretation of data (Rupp, 2002). That is, we explicitly build models to capture our substantive knowledge (Gelman et al., 1995). In educational assessment, the result is an incorporation of knowledge – be it about examinees, tasks, or the assessment context – into the statistical models used to guide inference (de Boeck & Wilson, 2004; Mislevy & Verhelst, 1990; Rupp & Mislevy, in press).

Relevant to the current work, the implication for model checking is clear. Techniques developed with a specific model, hypothesis, or assessment context in mind might not be easily repurposed to situations or settings that differ – even slightly – in their statistical or contextual nature. The primacy of unidimensional models in operational assessment has guided the development of model checking tools. However, tools explicitly built for assessing unidimensionality may not generalize to assess multidimensional models (Swaminathan et al., in press). A preferable model checking framework is one in which flexible tools may be re-instantiated in different situations, with each instance being tuned to the problem or desired inference at hand. We maintain that PPMC is such a framework and that it can be legitimately applied to criticize models, be they "off the shelf" or built in accordance with domain-specific knowledge.

To this end, this study examines PPMC for model criticism in light of inadequately accounting for the dimensional structure of observables in the context of multidimensional BNs with complex relations. The motivating example comes from cognitively-based psychometric modeling in computer networking, a rich, complex domain that often requires experts to bring multiple skills to bear in solving issues in the

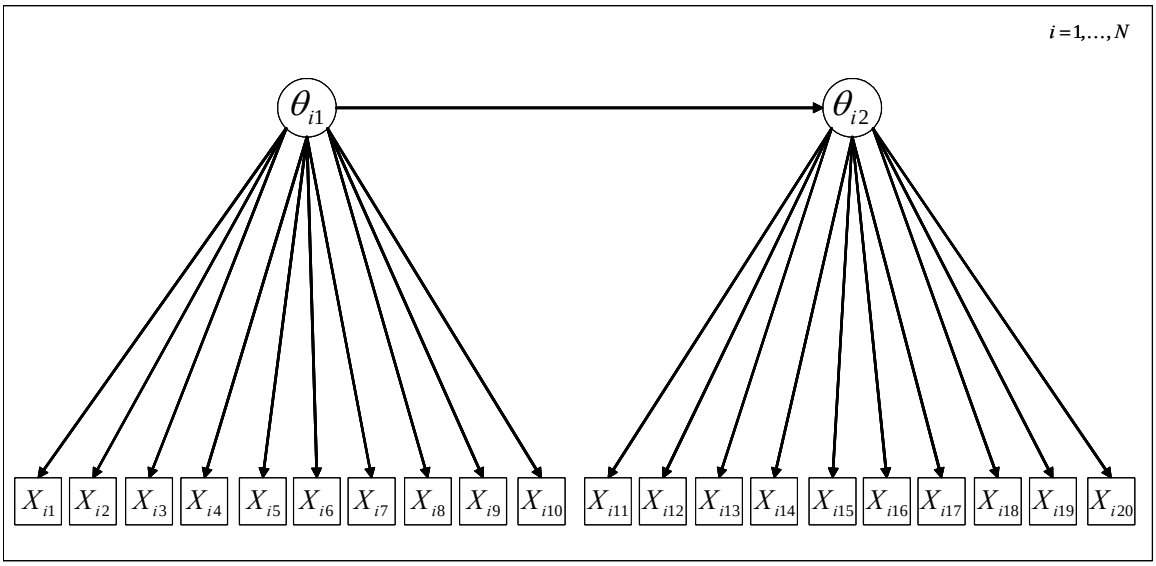
design, implementation, and troubleshooting of computer networks (Williamson et al., 2004). Accordingly, assessments of computer networking proficiency may be well served by (a) complex assessment tasks that require combinations of multiple focal skills and (b) their accompanying psychometric models (Behrens et al., 2004; Levy & Mislevy, 2004; Williamson et al., 2004).

BAYESIAN NETWORK MODELS

The two models considered below and the hypotheses they address are abstractions of the model for NetPASS, a simulation-based assessment of computer networking skills (Behrens et al., 2004). For simplicity, both models focus attention on an assessment consisting of 20 dichotomous observables directed at two skills that are captured by examinee variables labeled θ_1 and θ_2. As in NetPASS, the examinee variables are latent and discrete, taking on any of five values corresponding to the student's proficiency levels in terms of a four course sequence (Novice, Semester 1, Semester 2, Semester 3, Semester 4).

Figure 49 contains the directed graph for the first model. The plate surrounding the model indicates that the elements in the plate hold across examinees $i = 1, \ldots, N$. In referring to variables indexed by examinees (θ_{i1}, θ_{i2}, $X_{i1}, \ldots, X_{i20}$), we drop the subscript *i* when the expression, equation, or relation holds for all examinees.

Figure 49: Graphical BN model 1



The dependence between the skills is conveyed by the directed arrow from θ_1 to θ_2. The five levels of the examinee variables are coded as 1,…,5. With five levels for each of θ_1 and θ_2, the conditional probability table for θ_2 given θ_1 contains 25 cells. Direct estimation of so many probabilities (Spiegelhalter et al., 1993) is unwieldy. Principles from IRT and LCA offer a parsimonious way to model the conditional probabilities in large contingency tables (Formann, 1985; Levy & Mislevy, 2004; Mislevy et al., 2002).

The conditional probability of θ_2 given θ_1 is specified as following a constrained version of the graded response model (GRM; Samejima, 1969) via the effective theta method (Almond et al., 2001; Mislevy et al., 2002). The probability that θ_2 takes on a value of *k*, *k* = 1,…,5 is given as

$$P(\theta_2 = k \mid \theta_1) = P(\theta_2 \geq k \mid \theta_1) - P(\theta_2 \geq k+1 \mid \theta_1)$$

where

$$P(\theta_2 \geq k \mid \theta_1) = 1;\ k = 1$$

$$P(\theta_2 \geq k \mid \theta_1) = \text{logit}^{-1}(\theta_1^* - b_k);\ k = 2,\ldots,5$$

where the four b_k are category parameters and θ_1^* is the effective theta. Assuming the levels of θ_1 are equally spaced apart (an assumption which may be relaxed by estimating possibly unequal distances), we fix the values of b_k accordingly and define the effective theta θ_1^* via a linear function

$$\theta_1^* = c \times \theta_1 + d\ .$$

The slope (*c*) and intercept (*d*) parameters are akin to discrimination and difficulty parameters, respectively, in traditional IRT formulations. Note the simplicity of the model, there are two parameters to estimate, *c* and *d*, despite there being 25 cells in the five by five conditional probability table for θ_2 given θ_1.

The dichotomous observables are modeled using a Rasch model. More formally,

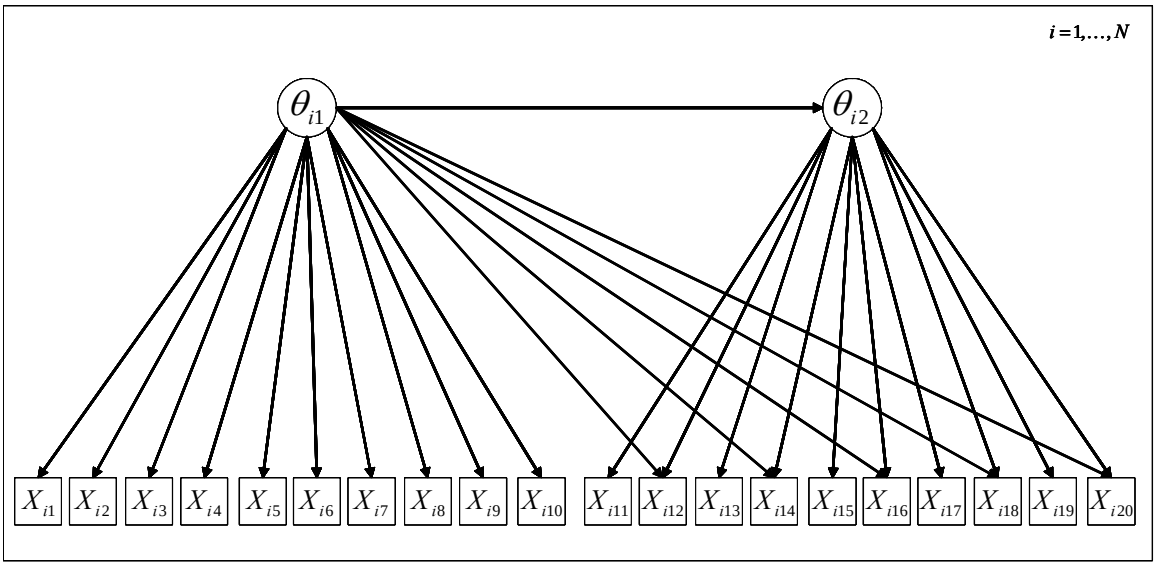
$$P(X_j = 1 \mid \theta_1) = \text{logit}^{-1}(\theta_1 - b_j);\ j = 1,\ldots,10$$

$$P(X_j = 1 \mid \theta_2) = \text{logit}^{-1}(\theta_2 - b_j);\ j = 11,\ldots,20$$

where b_j is the item difficulty parameter for observable *j*.

The second model alters the first by adding complexity to the skill set needed to successfully solve some of the tasks intended to inform upon θ_2. Figure 50 depicts this model, and differs from that in Figure 49 due to the addition of directed arrows from θ_1 to observables X_{12}, X_{14}, X_{16}, X_{18}, and X_{20}.

Figure 50: Graphical BN model 2



The form of the multidimensionality employed here is a variant on conjunctive MIRT, and explicitly models a hypothesis regarding the cognitive requirements for successful performance on computer networking tasks.

Suppose the dependency of performance on tasks on the skills is such that a certain level of θ_1 is needed, but after that threshold has been crossed, performance becomes primarily a function of θ_2. If task performance follows this structure, the model for the probability of observing a correct response to such a task is

$$P\left(X_j = 1 \mid \theta_1, \theta_2\right) = \text{logit}^{-1}\left(\theta_2 - b_j\right) \quad \text{if } \theta_1 > \delta_j \tag{19}$$

and zero otherwise, where δ_j is a threshold along the θ_1 dimension.

Following Almond et al. (2001), we term this an *inhibitory* relation, as possessing low levels of θ_1 inhibits performance. Upon passing some level (namely δ_j) of θ_1, the inhibition is lifted, and performance becomes a function of θ_2. As the conditioning

notation in Equation (19) expresses, observables following this model depend on multiple latent variables.

The models capture hypothesized relations among two skills in the domain and tasks built to assess those skills. Model 1 represents the idealized situation in which each task reflects one skill. Model 2 represents a situation in which performance on certain tasks depends on a complex (inhibitory) combination of multiple skills. One skill essentially serves as a prerequisite to performance on select tasks.

RESEARCH DESIGN

Data Generation

A brief simulation study was conducted to demonstrate the effectiveness of PPMC for performing model criticism. Three data sets of 1000 simulated examinees were generated from an inhibition model (model 2) and fit to a model that ignored the inhibitory relation (model 1). The distribution of θ_1 for all data sets is given in Table 5. Values for θ_2 were generated using the effective theta method with c = 1 and d = -0.5. The resulting conditional probabilities are given in Table 6. Note that the use of a negative value of d serves to render the expected value of θ_2 as slightly below that of θ_1, which models the situation in which the development of the second skill typically lags behind development of the first.

Table 5: Probability table for θ_1.

$P(\theta_1 = k)$				
1	2	3	4	5
.10	.20	.40	.20	.10

Table 6: Probability table for θ_2.

	$P(\theta_2 = k)$				
θ_1	1	2	3	4	5
1	.73	.15	.07	.03	.02
2	.50	.23	.15	.07	.05
3	.27	.23	.23	.15	.12
4	.12	.15	.23	.23	.27
5	.05	.07	.15	.23	.50

Item responses were generated via the inhibition model with IRT difficulty parameters given in Table 7. Note the symmetry within each set of 10 items around the value of 3.0. Thus the items are dispersed around the middle category of the latent variable (recall θ_1 and θ_2 are coded 1,…,5). By choosing items that span the difficulty continuum along θ_2 to be multidimensional, a possible confound between dimensionality and difficulty is avoided (Ackerman, 1996).

Table 7: Item parameters used in data generation.

Item	b_j	Item	b_j
1	1.0	11	1.0
2	1.5	12	1.0
3	2.0	13	2.0
4	2.5	14	2.0
5	3.0	15	3.0
6	3.0	16	3.0
7	3.5	17	4.0
8	4.0	18	4.0
9	4.5	19	5.0
10	5.0	20	5.0

The three data sets differed in the value of the threshold parameter δ_j for observables X_{12}, X_{14}, X_{16}, X_{18}, and X_{20}. The values for the δ_j used in generating the three data sets were 1.9, 2.9, and 3.9. These values serve to delimit which examinees are subject to inhibition. For $\delta_j = 1.9$, examinees with a value of one for θ_1 have zero probability of having a value of one for observables X_{12}, X_{14}, X_{16}, X_{18}, and X_{20}. Examinees with a value of 2 or larger for θ_1 have a probability of obtaining a value of one for each these observables; the probability for any one observable is defined by the Rasch model using the observable' b_j and the examinee's value for θ_2 (Equation (19)). For the data sets generated with δ_j values of 2.9 and 3.9, the necessary values of θ_1 are 3 and 4.[9] We refer to the conditions in which the δ_j are 1.9, 2.9, and 3.9 as *low*, *moderate*, and *high* inhibition conditions.

[9] Note that any value in the interval [1, 2) could have been used in place of 1.9, as θ_1 takes on integer values only. Likewise, any value in the intervals [2, 3) and [3, 4) could have been used in place of 2.9 and 3.9, respectively.

Model Estimation

For each data set, WinBUGS 1.4 (Spiegelhalter et al., 2003) was used to obtain an MCMC solution to model 1, which ignores the inhibitory relationship (code available from the author upon request). An assumption of exchangeability (de Finetti, 1964) with respect to the examinees allows for the specification of a common prior distribution for all examinee parameters (Lindley & Novick, 1981; Lindley & Smith, 1972). The probability distribution for θ_1 may be thought of as a multinomial distribution. As the natural conjugate for the multinomial, the Dirichlet distribution affords an interpretation of the parameters that govern the prior for the multinomial in terms of a sample size which may be compared to the size of the data set (Spiegelhalter et al., 1993). The values used for the Dirichlet distribution are $(1, 2, 4, 2, 1)$ and may be interpreted as expressing the prior belief that the probabilities in the multinomial distribution are the relative frequencies of the elements (i.e., $(0.1, 0.2, 0.4, 0.2, 0.1)$). The weight of this prior information is equal to the sum of the elements, 10, which is small in comparison to the data set of size 1000.

The priors for the parameters in the effective theta equation are

$$c \sim N(1, 10);$$

$$d \sim N(-.5, 10).$$

Turning to the observables, an assumption of exchangeability with respect to the tasks allows for the specification for a common prior for all item difficulty parameters:

$$b_j \sim N(3, 100).$$

The priors used are sufficiently diffuse in the chief regions of the distribution to allow the data to drive the solution.

Three chains from dispersed starting points were run for each analysis. Convergence was assessed via plots of the MCMC sequences. It was determined that the first 10000 iterations would be discarded as burn-in and the iterations used for PPMC should be thinned by a value of 10. A total of 11670 iterations were run for each of three chains. After discarding burn-in and thinning the chain, the resulting iterations were pooled to produce 501 iterations for use in PPMC.

PPMC

This simulation investigates features of the data at the level of the observables, rather than at the more global model level. The bivariate discrepancy measures investigated in the first study were also considered here.

RESULTS

We present and discuss the results of this study in different manner than was adopted in study 1. Here, the orientation will not be on typical behavior over repeated simulations, but rather on the ways PPMC may be conducted in practice to perform model criticism. We mimic the situation of an analyst in an applied setting by approaching and interpreting these results as if unaware of the true data-generating structure.

There are $(20 \times 19)/2 = 190$ nonredundant pairings of the 20 items. We leverage advances in representations of bivariate normal distributions (Murdoch & Chow, 1996) for visually displaying the PPP-values in the form of contour plots (Sinharay & Johnson, 2004). The contours trace the shape of a bivariate normal distribution with a given correlation. To transform the PPP-value to the metric of a correlation, each PPP-value is

multiplied by two and then subtracted by one. This value is then taken as the correlation in plotting the contours.

The result is that circular contours correspond to PPP-values of .5. Contours that are elongated and oriented positively correspond to positive PPP-values. For a maximal PPP-value of 1.0, the contour becomes a line with slope of 1. Conversely, contours that are elongated and oriented negatively correspond to negative PPP-values. For a minimal PPP-value of 0.0, the contour becomes a line with slope -1.

High Inhibition

Figures 51-58 contain the matrices of contours for the eight bivariate discrepancy measures. There are three different groups of results. The first group consists of the X^2 and G^2 discrepancy measures (Figures 51-52). These two are very similar to each other and are characterized by the absence of any contours that are strongly oriented in a positive direction. All the contours that deviate from near circularity are oriented in a negative direction, which corresponds to a low PPP-value. This is a manifestation of the X^2 and G^2 discrepancy measures insensitivity to direction of misfit (Chen & Thissen, 1997). Though they are able to detect presence of misfit, they do not speak to possible explanations or rectifications of the misfit.

Figure 51: Contours of PPP-values for X^2 for item-pairs in the high inhibition data.

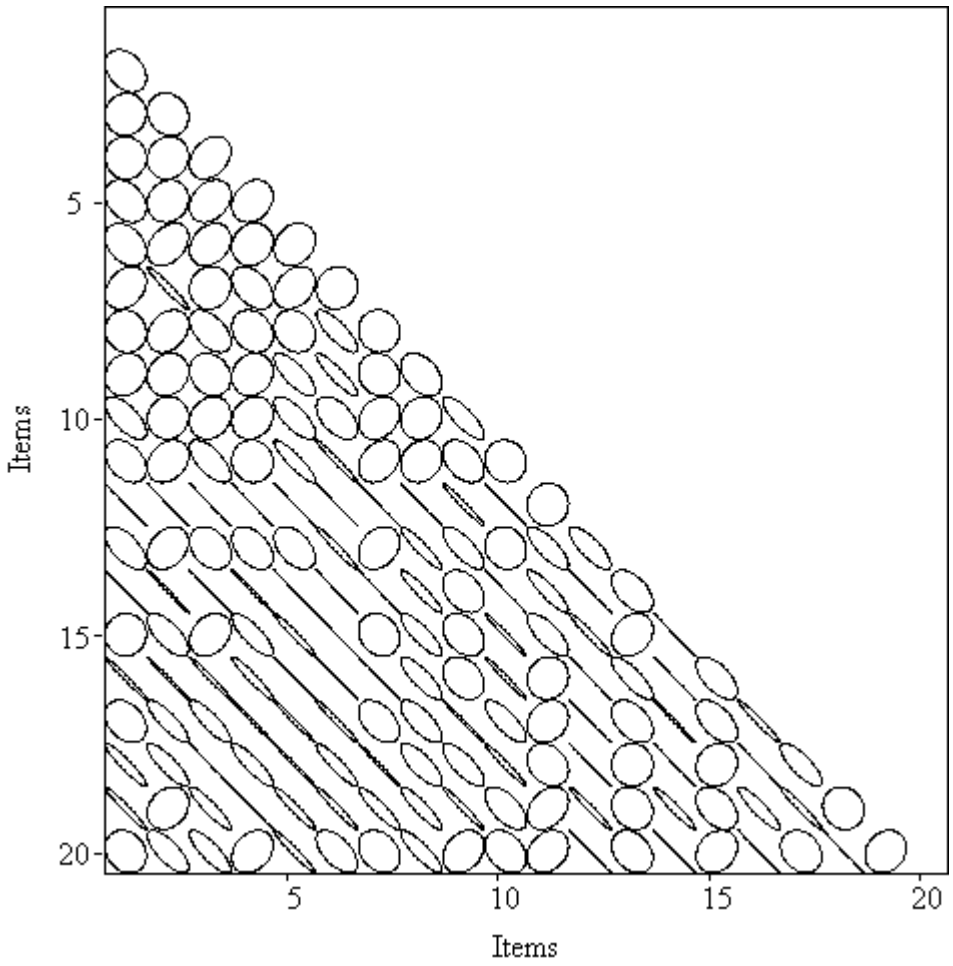
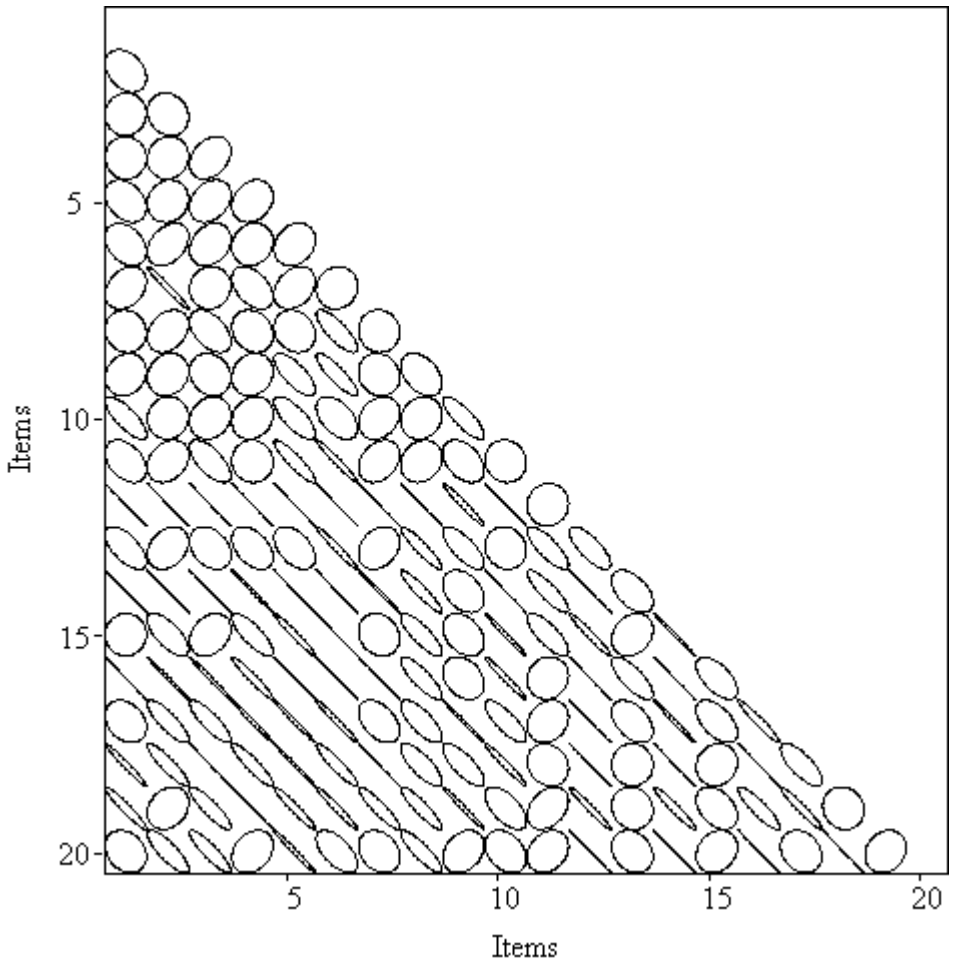


Figure 52: Contours of PPP-values for G^2 for item-pairs in the high inhibition data.



The second group is the model-based covariance and Q_3 (Figures 53-54). These are similar to each other and are distinguished by their pronounced pattern for the pairings of items 11-20. Though the remaining portions of the matrices of plots show some systematic behavior, the eye is drawn to the repeated 'X' pattern in the lower right part of the plots.

Figure 53: Contours of PPP-values for the model-based covariance for item-pairs in the high inhibition data.

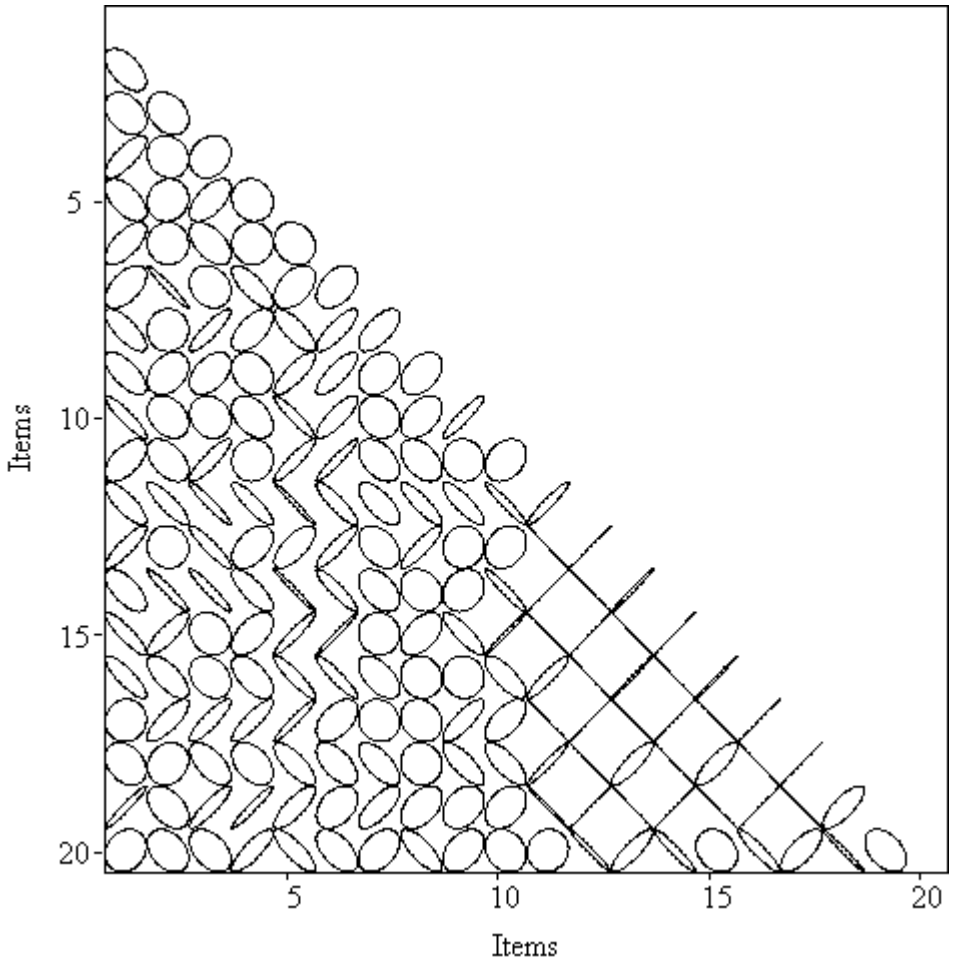
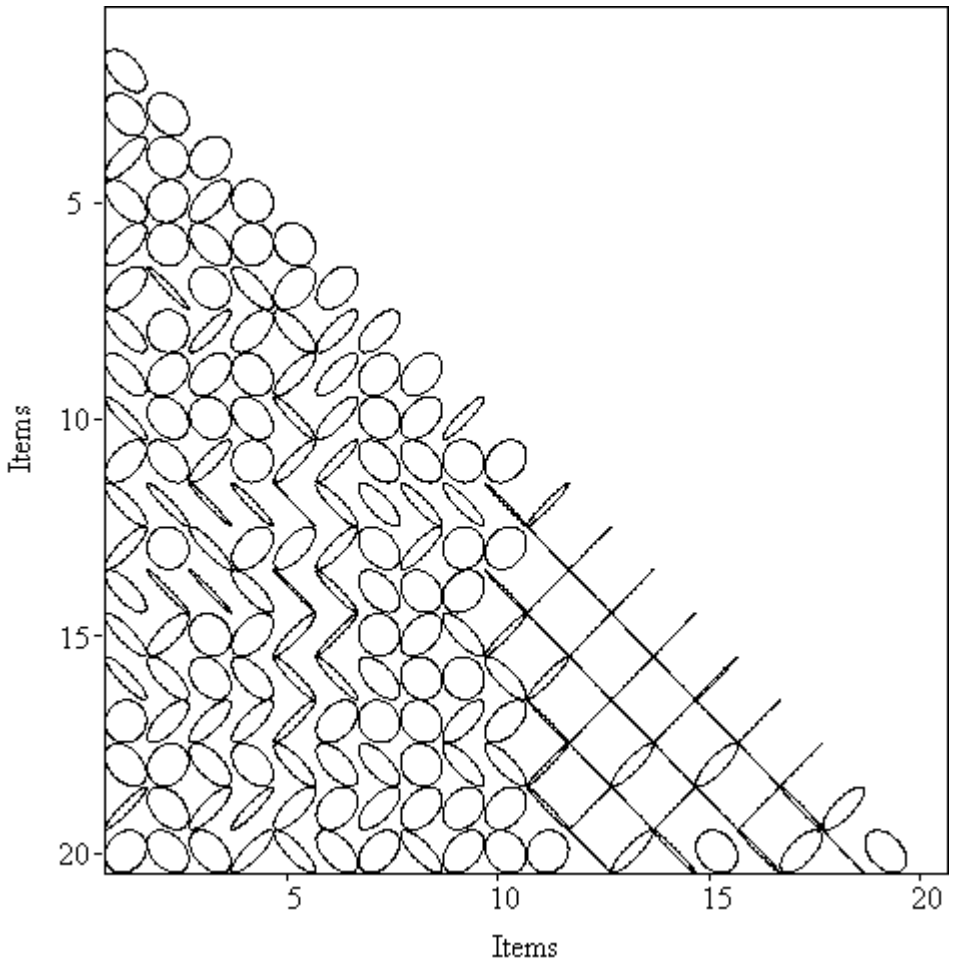


Figure 54: Contours of PPP-values for Q_3 for item-pairs in the high inhibition data.



The third group consists of the remaining discrepancy measures: the covariance, residual covariance, log odds ratio, and standardized log odds ratio residual (Figures 55-58). The results for all of these measures are similar to one another and are characterized by the sharp orientation of many of the contours for item-pairs with one element of the pair from the first 10 items and the second element of the pair from the second 10 items. Note also the absence of a crisp, repeated 'X' pattern in the lower right part of the plots; quite a few of the contours in this area are much more circular than their counterparts in Figures 53 and 54.

Figure 55: Contours of PPP-values for the covariance for item-pairs in the high inhibition data.

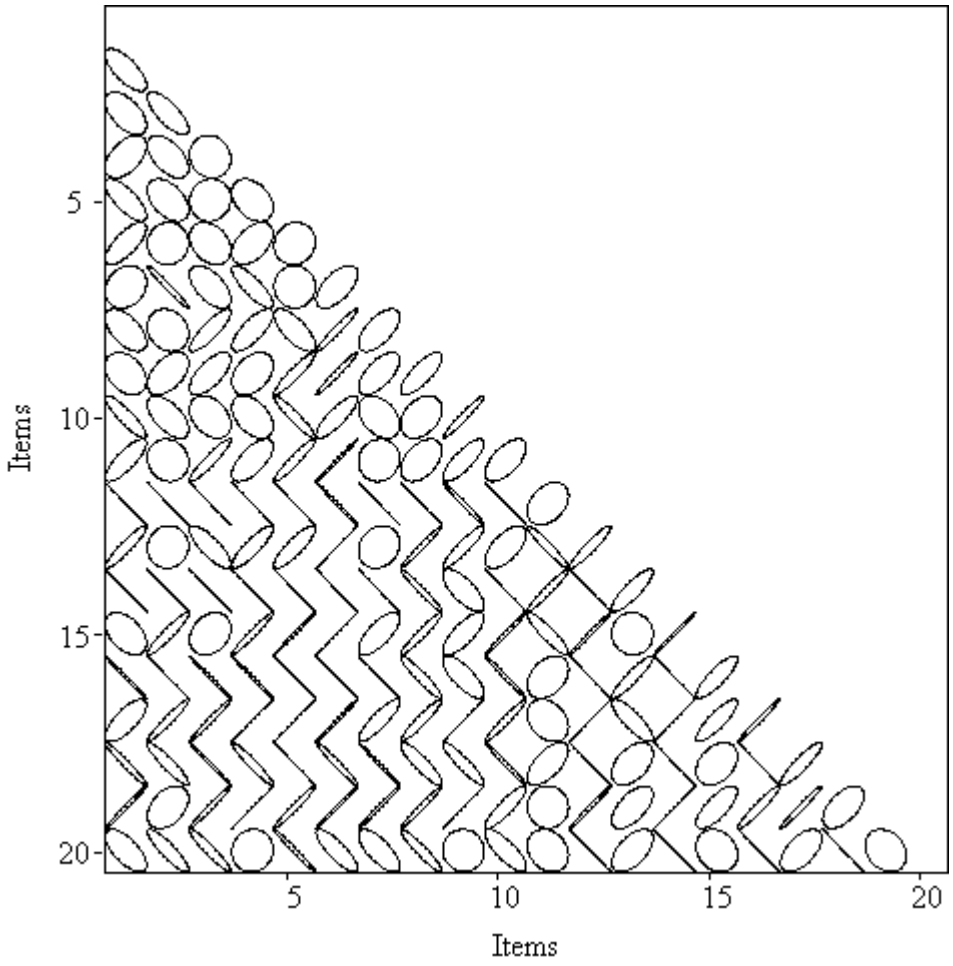


Figure 56: Contours of PPP-values for the log odds ratio for item-pairs in the high inhibition data.

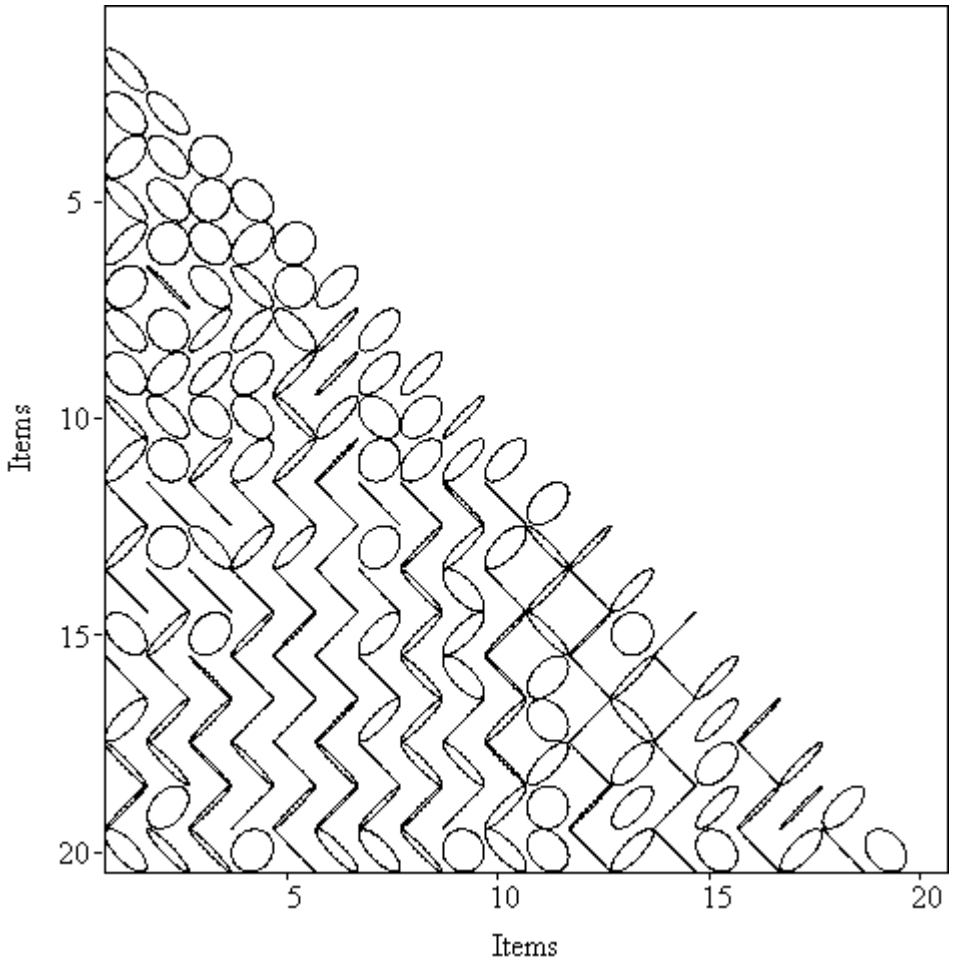


Figure 57: Contours of PPP-values for the residual covariance for item-pairs in the high inhibition data.

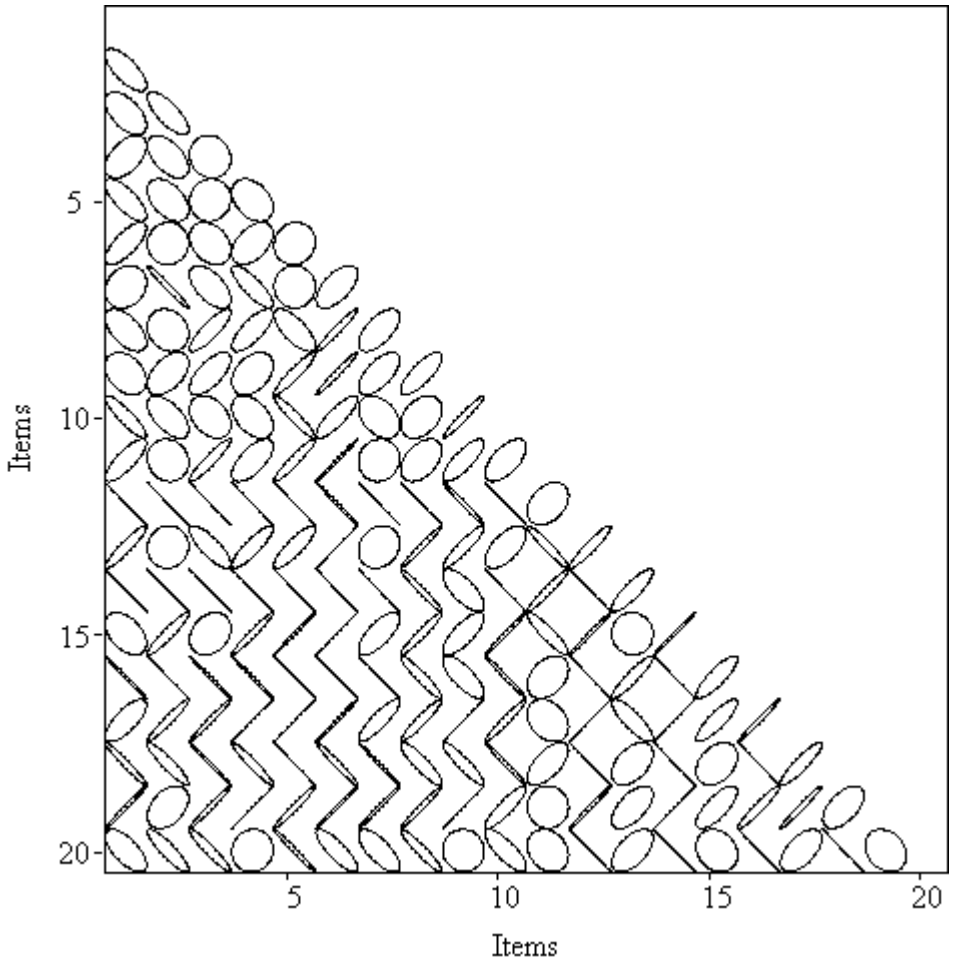
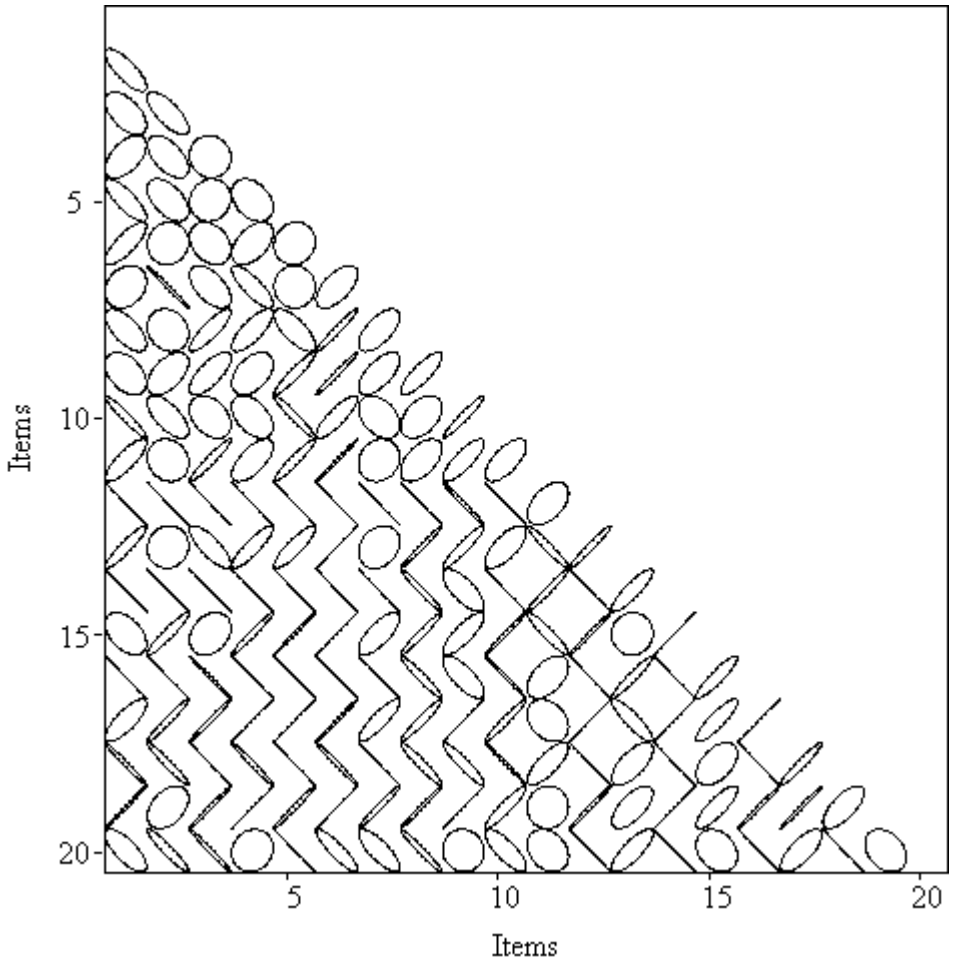


Figure 58: Contours of PPP-values for the standardized log odds ratio residual for item-pairs in the high inhibition data.



Analysis and Interpretation

We begin the analysis of these different groups of plots by starting with consensus. All of the discrepancy measures suggest that the pairings among items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} are sources of misfit. For the directional measures, the negatively oriented contours (and their associated PPP-values near 0) indicate that the measures for these item-pairs are being *underpredicted* by the model. Taken separately, this misfit in 10 pairings of these variables may be explained by the presence of as many

as 10 unmodeled sources of association. Taken collectively, that *all*10 pairings showed the same type of misfit suggests that the sources of misfit are most likely the same, or at least related. At this point, there is evidence that the model has failed to account for all sources of association among items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20}.

From here, there is some divergence regarding what items and item-pairs require attention. We take up a path based on the results of the model-based covariance and Q_3 (Figures 53-54). In addition to suggesting that an additional source of association among items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} is operating, the strongly negatively oriented contours suggest there are additional associations among items X_{11}, X_{13}, X_{15}, X_{17}, and X_{19}. What's more, the strongly *positively* oriented contours for the pairings of one of items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} with one of items X_{11}, X_{13}, X_{15}, X_{17}, and X_{19} reflect PPP-values close to 1. As discussed earlier, PPP-values close to 1 imply that the model is overpredicting the quantity under consideration. Thus, the model is *overpredicting* the associations in the pairings of one item from among X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} and one item from among X_{11}, X_{13}, X_{15}, X_{17}, and X_{19}. This suggests that the roots of the additional association among the first set of items are distinct from that of the second set.

To summarize, the lower right portions of Figures 53 and 54 suggest that (a) there may be additional sources of covariation for items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20}, (b) there may be additional sources of covariation for items X_{11}, X_{13}, X_{15}, X_{17}, and X_{19}, and (c) these possible unmodeled sources are distinct.

Reviewing the rest of the contours in these figures, patterns emerge for item-pairs in which one item is from the first half of the test and the other item is from the second half of the test. More specifically, the contours for the pairings of any of items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} with any of items $X_1,\ldots,X_{10}$ are, to varying degrees, negatively oriented. This implies the model is underpredicting the association among these variables. Conversely, the contours for the pairings of any of items X_{11}, X_{13}, X_{15}, X_{17}, and X_{19} with any of items $X_1,\ldots,X_{10}$ are positively oriented to some degree, which implies the model is overpredicting their associations.

Returning to the results for the covariance, residual covariance, log odds ratio, and standardized log odds ratio residual (Figures 55-58) these patterns for items on the first half of the test paired with items on the second half of the test are much sharper. However, the positively oriented contours for the pairings of one of items X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} with one of items X_{11}, X_{13}, X_{15}, X_{17}, and X_{19} are less sharp (i.e., more circular) than in the previous plots.

Both groups of directional discrepancy measures lead to similar portraits of the inadequacies of the model. By aggregating information across the discrepancy measures we are left with a strong impression that

- the model underpredicts associations among X_{12}, X_{14}, X_{16}, X_{18}, and X_{20};
- the model underpredicts the associations among X_{11}, X_{13}, X_{15}, X_{17}, and X_{19};
- the roots of the associations among X_{12}, X_{14}, X_{16}, X_{18}, and X_{20} are *not* same as those that underlie the associations among X_{11}, X_{13}, X_{15}, X_{17}, and X_{19};

- the model underpredicts the associations between any of X_{12}, X_{14}, X_{16}, X_{18}, or X_{20} and any of $X_1, \ldots, X_{10}$; and
- the model overpredicts the associations between any of items X_{11}, X_{13}, X_{15}, X_{17}, or X_{19} and any of items $X_1, \ldots, X_{10}$.

At this point, a number of statistically based model modifications may be considered. Though an argument could be constructed that implies the warranted modifications are those that would in fact result in the correct model, we do not pursue this in the current work. A more comprehensive approach to model modification would seek to incorporate the results of the PPMC with substantive considerations. The information from PPMC may be brought to bear with content experts, task authors, and other members of the assessment design team to better understand the shortcomings of the assumed model (Chen & Thissen, 1997; Sinharay, 2005; Zenisky et al., 2003).

Low and Moderate Inhibition

The patterns of the results for the low and moderate inhibition conditions mimicked those from the high inhibition condition. The only differences were observed in the magnitudes of the PPP-values. On the whole, the PPP-values were less extreme for the moderate inhibition compared to the high inhibition condition. Similarly, the PPP-values for the low inhibition condition were less extreme than the moderate inhibition condition. As a result, the contour plots for these conditions (not presented on space considerations) are slightly less sharply oriented. This is consistent with the findings in the first study. More specifically, in the conjunctive MIRT models, the difficulty along the auxiliary dimensions was manipulated to vary the strength of dependence; higher

values of the difficulty along the auxiliary dimensions resulted in more extreme PPP-values. In the current study, the inhibition parameter acts in a similar way, larger values indicate a stronger dependence on θ_1. Accordingly, larger values of the inhibition parameter resulted in more extreme PPP-values.

DISCUSSION AND CONCLUDING REMARKS

Two key points are brought to light from this study. First, PPMC can be a viable tool for conducting model criticism in light of concerns for inadequately accounting for multidimensionality across seemingly different settings. The situation examined here differs from that in the first study principally in the estimated model; here, the hypothesized model contains multiple latent variables. The PPMC and discrepancy measures employed in this research are therefore not merely an alternative method for assessing *unidimensionality*, but for assessing the model's specified dimensionality, *whatever that may be.* Swaminathan et al. (in press) noted that one limitation of popular confirmatory tools for dimensionality assessment is that they were principally designed to assess *unidimensionality*. As a flexible modeling tool, PPMC is easily adaptable to a variety of settings. This study has demonstrated that a number of existing tools for model criticism may be incorporated into a PPMC framework and applied to multidimensional models as well.

Second, this study serves to illustrate the benefits of considering *patterns* of results both across different instantiations of a discrepancy measure and across different discrepancy measures (Sinharay, 2005). Viewing PPP-values one by one may be efficient for descriptive purposes (e.g., identifying which item-pair has the most extreme

PPP-value for a certain discrepancy measure), but it is inefficient for supporting inferences more broadly conceived. Employing multiple discrepancy measures and tracking patterns, both within and between discrepancy measures, is potentially much more informative for substantive model criticism.

CHAPTER 6: CONCLUDING REMARKS

This work consists of two studies investigating the utility of PPMC for performing dimensionality assessment. The first pursued PPMC for criticizing unidimensional IRT models fit to data that follow compensatory or conjunctive MIRT models. 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 utility of PPMC – with appropriately chosen discrepancy measures – for performing dimensionality assessment was demonstrated.

In surveying the landscape of dimensionality assessment tools, Zhang and Stout (1999a) distinguished between parametric and non-parametric approaches. PPMC is parametric in the sense that an assumed model is fit, resulting in the posterior distribution. Beyond that, however, PPMC is quite flexible. In IRT, for example, nothing inherent in PPMC restricts it to be applied to logistic models studied here; PPMC may be applied when some other parametric form is assumed (e.g., normal-ogive models) or even when nonparametric item response models are employed (Karabatsos & Sheu, 2004).

In addition, Zhang and Stout (1999a) 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 it is not restricted to assess unidimensionality.

This last point was illustrated in the second study, which investigated PPMC in the context of multidimensional BNs. The key result is that PPMC may be employed to assess the model's specified dimensionality, whatever that may be. This underscores the general point that PPMC is a flexible framework for performing model criticism.

The primacy of unidimensional models used in the practice of assessment has guided the development of model checking tools. However, model checking tools explicitly built for assessing unidimensionality may not generalize easily to assess multidimensional models. It is argued that PPMC is a flexible family for model checking that can be instantiated in a variety of settings for a variety of models, be they familiar (e.g., unidimensional IRT) or specialized (e.g., multidimensional BN models with inhibitory relations).

Progressive approaches to psychometrics view statistical models as inferential tools to be built from components such that the end result is a structure for conducting probability-based reasoning that is grounded in substantive theory localized to the domain (Gelman et al., 1995; Mislevy et al., 2003; Pearl, 1988; Rupp, 2002). The flexibility of MCMC estimation for Bayesian models permits the application of models that hypothesize complex relations among observables and multiple latent variables, which pose difficulties for traditional approaches to model checking. The lack of adequate model checking tools might hinder the development and use of complex models rooted in substantive theory. As a framework for model assessment and criticism that is flexible enough to address complex models, PPMC may serve a valuable role in the movement towards the use innovative, theoretically based models. As such, PPMC may go beyond the limits of traditional statistical requirements for data-model fit assessment

and support efforts to bridge the divide between statistical modeling and substantive theory regarding the phenomena of interest.

Future work along these lines includes investigating PPMC for multidimensional models. That is, to what extent can PPMC inform upon model (in)adequacy when $M>1$ dimensions are hypothesized but $M^*>M$ dimensions underlie the data? Related future work might investigate the robustness of simple compensatory multidimensional models (possibly with interaction terms) against more complicated models such as those with conjunctive, disjunctive, or inhibitory relations.

Turning to the discrepancy measures themselves, it is argued that statistical and substantive consideration is necessary for selecting appropriate measures. A tradeoff for the flexibility of PPMC is that the onus is placed on the analyst to responsibly select the discrepancy measures. Summarizing the current studies, the model-implied covariance and Q_3 discrepancy measures performed the best. There is utility in considering multiple discrepancy measures in order to build an argument regarding the strengths and weaknesses of the model. Further research on all of the discrepancy measures is needed to fully explore their potential, particularly in consideration of discrepancy measures derived for related contexts (Hoijtink, 2001; Sinharay et al., in press).

Lastly, immediate extensions of this work include the application of PPMC to models for polytomous or continuous observables, as in graded IRT models or SEM. In particular, the model-implied covariance and Q_3 are deserving of future consideration. These measures were the most effective in the current studies and – in contrast to discrepancy measures explicitly built for dichotomous variables (e.g., odds ratios) – they ought to be easily instantiated for models of polytomous or continuous observables.

Developments in Bayesian modeling and estimation permit the construction and implementation of complex, substantively motivated statistical models. With the emergence of complex models comes the necessity of flexible model checking tools. The current work has evidenced the potential of PPMC to perform model criticism in terms of dimensionality assessment. As a research enterprise, PPMC is a rich area for practical application and methodological investigations. Further developments in PPMC are eagerly anticipated.

APPENDIX A: ON MCMC AND CONVERGENCE ASSESSMENT

The MCMC algorithm for fitting the 2-PL model employs a Metropolis-within-Gibbs algorithm (Patz & Junker, 1999a, 199b). In the following sections, this algorithm is described and the steps taken to determine the appropriate number of iterations necessary for convergence are reported.

GIBBS SAMPLING

Let $\Omega_1,\ldots,\Omega_R$ be the *R* components of the unknown parameters contained in $\boldsymbol{\Omega}$. The desired distribution is the posterior distribution of the unknown parameters after observing the data, $P(\boldsymbol{\Omega}\,|\,\mathbf{X})$, which may be defined by the complete set of such *full conditional* distributions (Gelfand & Smith, 1990). A full conditional distribution for a variable is its distribution conditional on all remaining variables. The full conditionals are then

$$
\begin{aligned}
&P(\Omega_1\,|\,\Omega_2,\Omega_3,\ldots,\Omega_R,\mathbf{X});\\
&P(\Omega_2\,|\,\Omega_1,\Omega_3,\ldots,\Omega_R,\mathbf{X});\\
&\vdots\\
&P(\Omega_r\,|\,\Omega_1,\ldots,\Omega_{r-1},\Omega_{r+1},\ldots\Omega_R,\mathbf{X});\\
&\vdots\\
&P(\Omega_R\,|\,\Omega_1,\Omega_2,\ldots,\Omega_{R-1},\mathbf{X}).
\end{aligned}
$$

A full conditional may be evaluated via Bayes' theorem. For any parameter Ω_r,

$$
\begin{aligned}
&P(\Omega_r\,|\,\Omega_1,\ldots,\Omega_{r-1},\Omega_{r+1},\ldots\Omega_R,\mathbf{X})\\
&\qquad = \frac{P(\Omega_r)\times P(\Omega_1,\ldots,\Omega_{r-1},\Omega_{r+1},\ldots\Omega_R,\mathbf{X}\,|\,\Omega_r)}{\int P(\Omega_r)\times P(\Omega_1,\ldots,\Omega_{r-1},\Omega_{r+1},\ldots\Omega_R,\mathbf{X}\,|\,\Omega_r)\,d\Omega_r} \qquad \text{(A1)}\\
&\qquad \propto P(\Omega_r)\times P(\Omega_1,\ldots,\Omega_{r-1},\Omega_{r+1},\ldots\Omega_R,\mathbf{X}\,|\,\Omega_r)
\end{aligned}
$$

The first term on the right of Equation (A1) is the prior distribution for Ω_r. The second term is the conditional probability of the remaining variables given Ω_r. Conditional independence assumptions may be invoked to eliminate terms from this latter expression, thereby reducing the computational complexity.

A Gibbs sampling scheme (Casella & George, 1992; Gelfand & Smith, 1990; Gilks et al., 1996a) proceeds by initializing the parameters as $\Omega_1^0, \ldots, \Omega_R^0$ and then iteratively sampling from the full conditionals where the values of remaining unknown parameters are set at their current values. The first iteration in Gibbs sampler performs the following draws:

$$
\begin{aligned}
&\Omega_1^1 \sim P\left(\Omega_1 \mid \Omega_2^0, \Omega_3^0, \ldots, \Omega_R^0, \mathbf{X}\right);\\
&\Omega_2^1 \sim P\left(\Omega_1 \mid \Omega_1^1, \Omega_3^0, \ldots, \Omega_R^0, \mathbf{X}\right);\\
&\vdots\\
&\Omega_r^1 \sim P\left(\Omega_r \mid \Omega_1^1, \ldots, \Omega_{r-1}^1, \Omega_{r+1}^0, \ldots, \Omega_R^0, \mathbf{X}\right);\\
&\vdots\\
&\Omega_R^1 \sim P\left(\Omega_R \mid \Omega_1^1, \Omega_2^1, \ldots, \Omega_{R-1}^1, \mathbf{X}\right).
\end{aligned}
$$

The collection $\Omega_1^1, \ldots, \Omega_R^1$ constitutes the first iteration of a Gibbs cycle. This process is repeated for *K* iterations, where the draw for parameter *r* at iteration *k* may be expressed as

$$\Omega_r^k \sim P\left(\Omega_r \mid \Omega_{<r}^k, \Omega_{>r}^{k-1}, \mathbf{X}\right).$$

Metropolis-Within-Gibbs

The full conditionals for the 2-PL model are not tractable (Maris & Bechger, in preparation). As a consequence, draws for the parameters in the Gibbs cycles cannot be obtained by directly sampling from the full conditional distributions. To enable sampling from each full conditional, a Metropolis step is employed at each stage of the Gibbs

sampler (Chib & Greenberg, 1995; Patz & Junker, 1999a, 1999b). The Metropolis algorithm is briefly reviewed next. More complete details are given by Chib and Greenberg (1995).

Let π be the target distribution of interest to be sampled from. In terms of Gibbs sampling just described, π is the full conditional for the parameter. The Metropolis algorithm takes a random draw *y* from a *proposal* distribution, *q*. The algorithm requires that *q* be symmetric with respect to its arguments (see Chib & Greenberg, 1995 for an overview of Metropolis-Hastings sampling, which relaxes this restriction). A popular choice for *q* is the normal distribution, which is symmetric with respect to the value and the mean of the distribution, as the height of the normal pdf is unchanged if *x* and μ reverse roles: $N\left(x \mid \mu, \sigma^2\right) = N\left(\mu \mid x, \sigma^2\right)$ (Chib & Greenberg, 1995). The sampler accepts this candidate point as the next value for the parameter with probability defined by the ratio of the heights of the candidate and current point in the posterior distribution (Chib & Greenberg, 1995). If the candidate is not accepted, the current value is retained as the next value in the chain.

More formally, for any parameter Ω_r we draw *y* via

$$y \sim q\left(y \mid \Omega_r^k\right). \tag{A2}$$

The conditioning notation in Equation (A2) expresses that the proposal distribution *q* may be dependent on the current value for the parameter, Ω_r^k. In estimating the 2-PL in the current work, a current point Metropolis sampler is used in which $q\left(\bullet \mid \Omega_r^k\right)$ is a normal distribution with mean Ω_r^k.

An acceptance probability is calculated as

$$\alpha = \min\left[1, \frac{\pi(y)}{\pi(\Omega_r^k)}\right].$$

The candidate point *y* is accepted as the value for the next iteration, Ω_r^{k+1}, with probability α and the current point Ω_r^k is retained as the value for Ω_r^{k+1} with probability $1-\alpha$.

In sum, the MCMC estimation proceeds by repeating a large number of iterations of the Gibbs sampler for the unknown parameters. In every iteration, the value for each parameter is obtained by performing a Metropolis step, using a normal proposal distribution centered at the current point.

CONVERGENCE ASSESSMENT

The choice of the standard deviation of the proposal distribution is crucial for how the chain mixes, converges, and moves around the posterior. An ideal situation would be to vary the proposal distributions (in terms of the standard deviation) for each data set. Owing to the size of the study (N+2J parameters per chain, 5 chains per replication, 50 replications per cell, 291 cells) such an approach is impractical. Instead, a pilot study was conducted to determine a value for the proposal distributions' standard deviations and the requisite number of iterations necessary to achieve convergence.

An optimal or near optimal choice to obtain adequate mixing is to set the standard deviation of the proposal distribution equal to 2.4 multiplied by the posterior standard deviation (Gelman et al., 1995). A number of alternatives for the standard deviation were tried for several pilot data sets. The results of these runs led to comparable estimates of

the posterior means and variances. That is, regardless of the choice of the standard deviation of the proposal distribution, all estimation routines converged to the same distributions. This result is one of the most appealing properties of MCMC estimation: under mild regularity conditions, the chain will converge to stationarity regardless of the proposal distribution (Smith & Roberts, 1993).

However, varying the standard deviation of the proposal distribution *did* have considerable impact on the mixing rates and the number of iterations necessary for convergence. As the guideline above (2.4 multiplied by the posterior standard deviation) depends on the posterior standard deviation, it is sensible to expect that different results would be obtained for data sets of different sample sizes, as the posterior variability for item parameters is much smaller when sample size is 2500 than when it is 750 or 250. This is exactly what was observed.

After the pilot runs were used to assess the posterior variability of the item parameters in the different sample sizes, the 2.4 multiplied by the posterior standard deviation guideline was applied to yield the near optimal values. Based on these analyses, a standard deviation of .5 was used for the proposal distributions for item parameters in the data sets with a sample size of 250. A standard deviation of .3 was used for data sets with a sample size of 750. Lastly, a standard deviation of .15 was used for data sets with a sample size of 250. These values resulted in near optimal combinations of adequate acceptance rates and fast convergence.

Convergence assessment using these values for the standard deviations for the proposal distributions was conducted for other cells in the study to assess whether alternative proposal distributions and/or an alternative number of iterations were

necessary for calibrating the unidimensional model to multidimensional data. Though not discussed on space considerations, these pilot analyses revealed no differences in mixing rate or number of iterations necessary for convergence. This has two implications. For the IRT study, sample size specific proposal distributions and burn-in criteria are warranted for all estimation runs and do not need to be adjusted for the different combinations of the manipulated factors. More broadly, this suggests that convergence need not imply model adequacy. That is, fast convergence and adequate mixing is not evidence that the model is adequate. However, the possibility remains open that failure to converge may constitute evidence of model inadequacy.

APPENDIX B: OBTAINING EXPECTED COUNTS OF FREQUENCIES

Here we take up the issue of ways to obtain expected frequencies, as are required in the computation of a number of discrepancy measures. We confine our attention to obtaining the expected frequency for a single variable. The logic may be easily extended to obtaining expected counts of joint frequencies as needed.

Let $n_j = \sum_{i=1}^{N} X_{ij}$ be the number correct for item *j*. Several ways to formulate the expected value of this quantity are discussed below. One argument proceeds as follows. The expected number correct is just the number of examinees multiplied by the model-implied probability of answering the item correctly:

$$E(n_j) = N \times P(X_j = 1 \mid b_j, a_j). \tag{B1}$$

But IRT models, including those studied in this work, typically do not specify the *marginal* probability of correct response, but the *conditional* probability given θ. To get the marginal probability, one needs to eliminate the dependence on θ via marginalization. Because θ is continuous, this comes to integrating over θ:

$$P(X_j = 1 \mid b_j, a_j) = \int_{\theta} P(X_j = 1 \mid \theta, b_j, a_j) P(\theta) d\theta . \tag{B2}$$

In practice, numerical integration may be used to approximate this integral. The integral is replaced by the sum over quadrature points

$$\int_{\theta} P(X_j = 1 \mid \theta, b_j, a_j) P(\theta) d\theta \approx \sum_{q} P(X_j = 1 \mid \theta_q, b_j, a_j) w(q) \tag{B3}$$

where θ_q is the value along the latent continuum of point *q* and $w(q)$ is the height of point in the distribution at θ_q. The $w(q)$ serve to weight the (conditional) probabilities. A common choice for the quadrature points is to employ 41 equally spaced points from

-4 to $+4$, which spans the latent continuum where the majority of the examinees are located (assuming appropriate scaling choices).

The key question in this line of reasoning is then: what is the distribution of θ? In marginal maximum likelihood (MML), $P(\theta)$ is interpreted as a (prior) population distribution that is marginalized over. It is often assumed to be $N(0,1)$, which also serves to identify the scale. In a Bayesian analysis, to obtain the expected frequency based on the calibration of the model requires the use of the posterior distribution, $P(\theta \mid \mathbf{X})$. How might this be done via MCMC, which uses draws from the posterior? Three ways come to mind.

Draw *k* from the posterior results in θ_i^k, the value of θ for examinee *i* from iteration *k*, for *i* = 1,…,*N*. One way to numerically approximate the posterior is to assume it is normal and calculate the mean and variance of the θ_i^k and use these as the mean and variance of assumed normal distribution. This defines a particular normal distribution, and quadrature points can be defined as usual. This constitutes a fast, though rough, approximation.

A second way is to sort the θ_i^k into a large number, say 50, bins. For each bin, the mean of the θ_i^k can be calculated as can the relative frequency of the bin. The means of the bins could be used as the quadrature points with the weights being the relative frequency. Another way to bin the draws would be to set out the bins, say 42 of them based around the familiar 41 quadrature points from -4 to +4. The θ_i^k can then be sorted into these bins and the relative frequencies can be calculated to serve as the weights in the numerical integration.

The third way assumes neither the form of the posterior nor the (somewhat arbitrary) definition of bins. This way employs values of θ_i^k directly. This treats the N values of θ_i^k as the quadrature points, each with weight 1/N. Following this formulation the approximation becomes

$$\int_{\theta} P\left(X_j = 1 \mid \theta, b_j, a_j\right) P(\theta) d\theta \approx \sum_i P\left(X_{ij} = 1 \mid \theta_i, b_j, a_j\right) 1/N. \tag{B4}$$

Relative to Equation (B3), note the change in the subscript of θ from q to i and the addition of the subscript i to X; accordingly the sum is over i. Substituting in for the equation for the expected number correct:

$$\begin{aligned} E\left(n_j\right) &= N \times P\left(X_j = 1 \mid b_j, a_j\right) \\ &\approx N \times \sum_i P\left(X_{ij} = 1 \mid \theta_i, b_j, a_j\right) 1/N. \\ &= \sum_i P\left(X_{ij} = 1 \mid \theta_i, b_j, a_j\right) \end{aligned} \tag{B5}$$

Under this formulation, the expected number correct for an item is the sum over examinees of the IRT model-implied probability of correct for each examinee for that item.

The above argument was based on the reasoning that the expected number correct could be thought of as the sample size N multiplied by the (marginal) probability of a correct response (Equation (B1)). A different argument proceeds as follows. Again, let $n_j = \sum_{i=1}^{N} X_{ij}$. The expected value of this quantity may be formulated as

$$E\left(n_j\right) = E\left(\sum_i X_{ij}\right) = \sum_i E\left(X_{ij}\right) = \sum_i P\left(X_{ij} = 1 \mid \theta_i, b_j, a_j\right). \tag{B6}$$

Note that this is the exact same result as was arrived at under the previous argument treating the draws for θ as quadrature points (equation (B5)). Two differences come to

mind. The first is that in the first derivation of this formula, the sum of the probabilities over examinees is regarded as an *approximation* to the expected number correct while in the second there is no expression of it being an approximation. The difference between these two interpretations is in the desired inference. In the first case, the N values of θ are thought of as drawn from a *population* of θ's. In the second case, the N values of θ are treated separately as their own entity.

This is connected to the way that the prior and posterior distributions are interpreted. In the first approach, the (posterior) distribution of θ is thought of as a distribution *for the population*. The second approach is more in line with thinking of the (posterior) distribution of θ as a distribution *for the examinee, one examinee at a time*. A related question comes from how the prior $P(\theta)$ is thought of in MML vs. fully Bayesian approaches. In MML, it is regarded as a population distribution. In the fully Bayesian approach, we specify $P(\theta)$ as the prior distribution *for each examinee's* θ_i. Typically, we specify the same prior for each examinee as may be warranted by an assumption of exchangeability (Lindley & Novick, 1981; Lindley & Smith, 1972).

A larger treatment of these issues is beyond the current scope. Further work is needed in understanding the various alternative conceptualizations of probability expressions, and their implications – if any – for obtaining expected frequencies in Bayesian modeling and PPMC.

APPENDIX C: ON THE EXCHANGEABILITY ASSUMPTIONS

EXCHANGEABILITY ASSUMPTIONS REGARDING AUXILIARY DIMENSIONS

In the analyses of the univariate discrepancy measures, results for items reflecting θ_1 and θ_2 were pooled with items reflecting θ_1 and θ_3. Similarly, in the analyses of the bivariate discrepancy measures, item-pairs in which both items reflect θ_1 and θ_2 were pooled with item-pairs in which both items reflect θ_1 and θ_3 (and referred to as item-pairs in which both items reflect the same multiple dimensions). Likewise, item-pairs in which one item reflected θ_1 and the other reflected θ_1 and θ_2 were pooled with item-pairs in which one item reflected θ_1 and the other reflected θ_1 and θ_3 (and referred to as item-pairs in which one item reflects the primary dimension only and the other reflects multiple dimensions).

These choices regarding what to pool reflect explicit exchangeability assumptions (de Finetti, 1964). There is reason to believe (indeed, hope) that PPMC will lead to different results for items that reflect multiple dimensions as opposed to items that only reflect the primary dimension; hence the use of two sets of points. However, there is no reason to believe that PPMC will lead to different results for items that reflect θ_1 and θ_2 as compared to items that reflect θ_1 and θ_3. As such, items that reflected θ_1 and θ_2 and items that reflected θ_1 and θ_3 were assumed to be exchangeable.

The foundation of this assumption is the balancing of the research design with respect to the second and third dimension. In each analysis, (a) the same number of items reflect θ_2 as reflect θ_3, (b) the strength of dependence on θ_2 and θ_3 are equal, (c) the

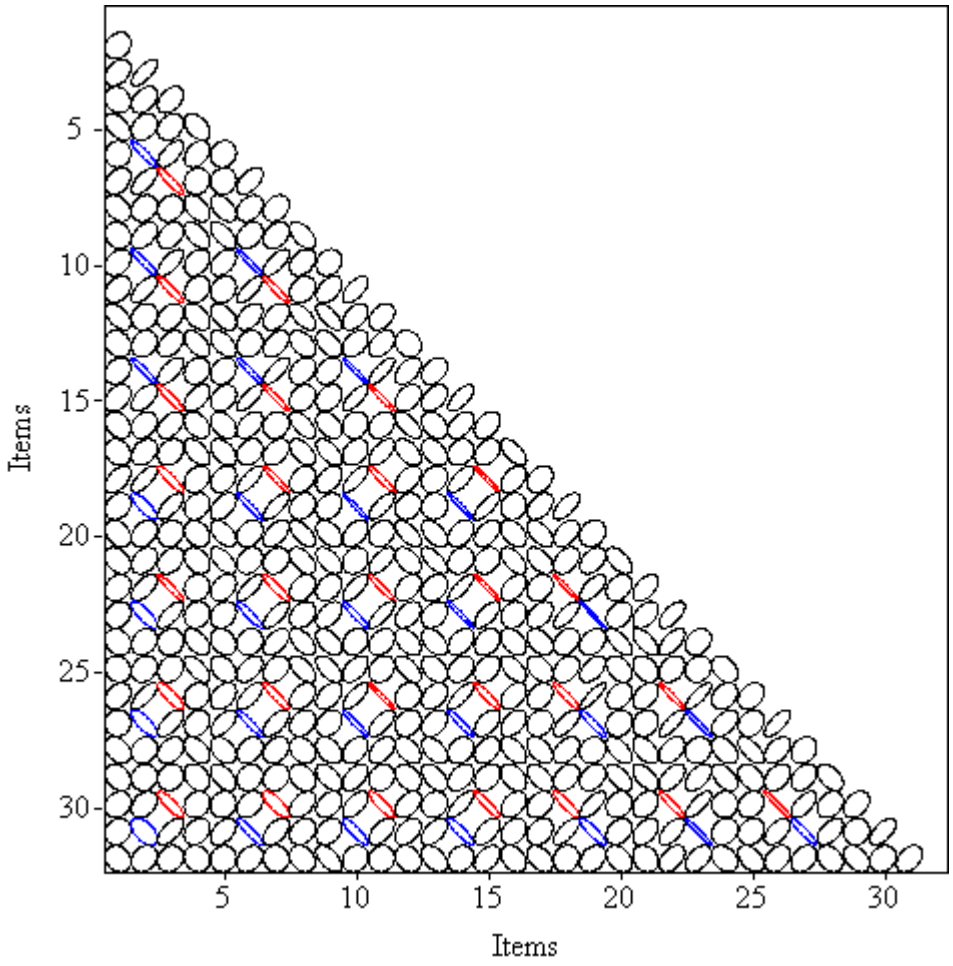
correlation between θ_1 and θ_2 is the same as the correlation between θ_1 and θ_3, and (d) the distribution of item difficulties for items that reflect θ_1 and θ_2 is comparable to the distribution of item difficulties for items that reflect θ_1 and θ_3. In the following sections, this assumption is evaluated.

COMPENSATORY MULTIDIMENSIONAL DATA

The first step taken was to examine contour plots for the median PPP-values for each discrepancy measure in each condition. Figure C1 plots these contours for the case for compensatory MIRT where the a_{j2} and a_{j3} were .75, the correlations between the dimensions were 0.3 and 16 of the 32 items reflected multiple dimensions. The plots are arranged as the lower triangle of a matrix of pairings of items. Each element of the matrix corresponds to the pairing of the items that define the row and column, respectively.

The plots were developed following advances in protocols for representing the association in bivariate normal distributions (Murdoch & Chow, 1996; Sinharay & Johnson, 2004). Circular contours correspond to PPP-values of .5. Contours that are elongated and oriented positively correspond to positive PPP-values. For the maximal PPP-value of 1.0, the contour becomes a line with slope of 1. Conversely, contours that are elongated and oriented negatively correspond to negative PPP-values. For the minimal PPP-value of 0.0, the contour becomes a line with slope -1.

Figure C1: Contour plots of median PPP-values disaggregated by dimension based on conducting PPMC on compensatory multidimensional data.



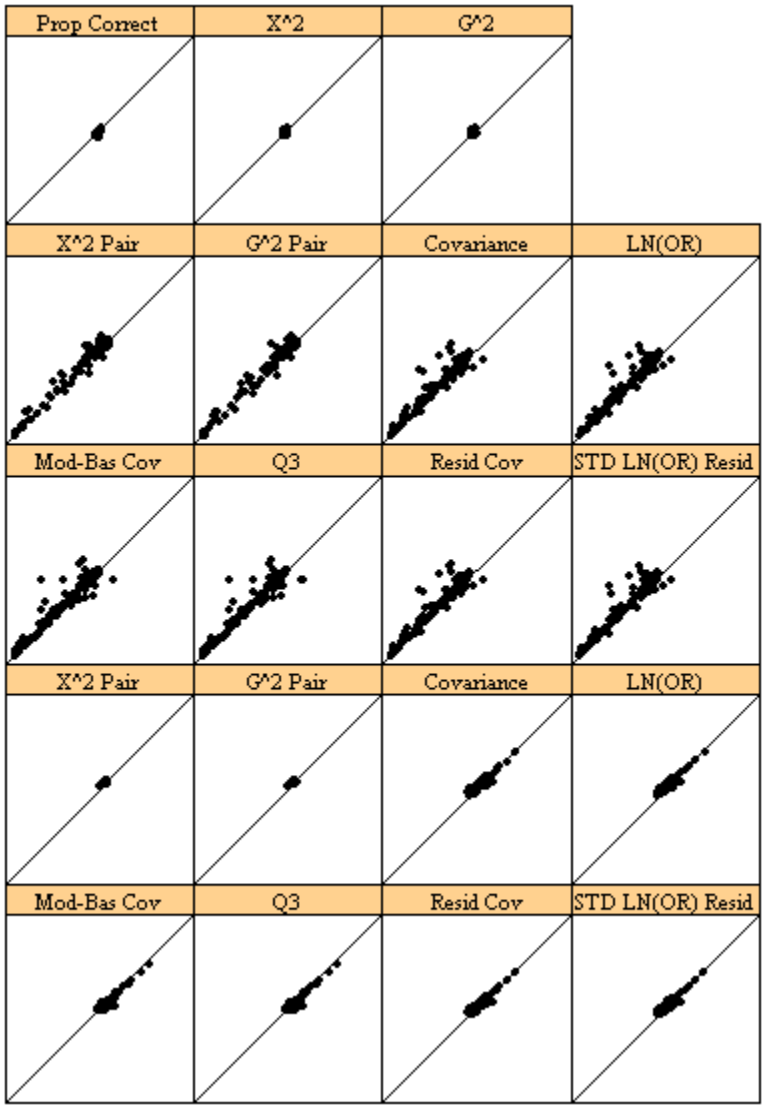
In Figure C1, item-pairs in which both items reflect θ_1 and θ_2 are plotted in red. Item-pairs in which both items reflect θ_1 and θ_3 are plotted in blue. First, note that all of these contours are elongated and negatively oriented, indicative of low PPP-values. Of interest in this analysis is the comparison between items that reflect θ_2 and items that reflect θ_3. No systematic differences between the red and blue contours are evident.

On space considerations, contour plots for all the discrepancy measures in all the conditions are not presented. To more comprehensively assess the appropriateness of the

exchangeability, the follow procedure was adopted. For univariate discrepancy measures, in each condition the median PPP-value across all replications and all items that reflect θ_1 and θ_2 was calculated. Similarly, the median PPP-value across all replications and all items that reflect θ_1 and θ_3 was calculated. In other words, items reflecting θ_1 and θ_2 were pooled with each other, and items reflecting θ_1 and θ_3 were pooled with each other, but the two groups were kept separate.

The first row in Figure C2 contains panels with scatterplots of the medians for the three univariate discrepancy measures for the analysis of compensatory multidimensional data. Each condition (i.e., combination of strength of dependence, correlations among the dimensions, proportion of multidimensional items, and sample size) is a single point; the value along the horizontal axis corresponds to the median PPP-value for the discrepancy measure for items that reflect θ_1 and θ_2 and the value along the vertical axis corresponds to the median PPP-value for the discrepancy measure for items that reflect θ_1 and θ_3.

Figure C2: Scatterplots of median PPP-values disaggregated by dimension based on conducting PPMC on compensatory multidimensional data.



An analogous approach was taken in analyzing the bivariate discrepancy measures. For each measure and each condition there are two plots rather than one. The first of these scatterplots contains points defined by:

- the median PPP-value evaluated on item-pairs that measure θ_1 and θ_2 (horizontal axis) with

- the median PPP-value evaluated on item-pairs that measure θ_1 and θ_3 (vertical axis).

These plots for the bivariate discrepancy measures based on the compensatory multidimensional data are contained in the second and third rows of Figure C2. The second plot for each bivariate discrepancy measure contains points defined by

- the median PPP-value evaluated on item-pairs that in which the first item reflects θ_1 only and the second item reflects θ_1 and θ_2 (horizontal axis) with
- the median PPP-value evaluated on item-pairs that in which the first item reflects θ_1 only and the second item reflects θ_1 and θ_3 (vertical axis).

These plots for the bivariate discrepancy measures based on the compensatory multidimensional data are contained in the fourth and fifth rows of Figure C2. Relative to the pooled analyses, these plots disaggregate the item-pairs in which one item reflects θ_1 only and the second item reflects multiple dimensions.

In each plot in Figure C2, the solid line is not a regression line, but rather the unit line denoting equality between the median PPP-values that reflect θ_1 and θ_2 with those that reflect θ_1 and θ_3. With only slight deviations, all the plots contain points that are randomly scattered around the unit line. This indicates that there is no reason to suggest that either the items reflecting θ_2 or those reflecting θ_3 will lead to larger PPP-values than the other. In most plots, the points are tightly scattered around the unit line, indicating that there is little variation between the PPP-values for items that reflect θ_1 and θ_2 and those items that reflect θ_1 and θ_3.

Substantively, there is strong evidence to suggest that the exchangeability assumptions regarding items that reflect θ_2 and items that reflect θ_3 are justified. This supports the pooling of items and item-pairs from the two auxiliary dimensions.

CONJUNCTIVE MULTIDIMENSIONAL DATA

Figure C3 contains scatterplots of the medians for the discrepancy measures for the analysis of conjunctive multidimensional data. Again, the unit line is superimposed and is not a regression line. As was the case of compensatory multidimensional data, in each plot the points appear randomly but tightly distributed around the unit line. The PPP-values for items that reflect θ_1 and θ_2 do not vary systematically or substantially from the PPP-values for items that reflect θ_1 and θ_3.

An analysis of the contour plots for conjunctive MIRT data reveals an unexpected finding. To illustrate, Figure C4 contains the contour plots for the conjunctive MIRT data where the difficulty along the auxiliary dimensions is -0.5, the correlations between the dimensions is 0.3, and 16 of the 32 items reflect multiple dimensions. For contours for multidimensional item-pairs that are close to one another, the shapes are very similar, which supports the exchangeability assumption regarding items that reflect θ_1 and θ_2 and the items that reflect θ_1 and θ_3.

Figure C3: Scatterplots of median PPP-values disaggregated by dimension based on conducting PPMC on conjunctive multidimensional data.

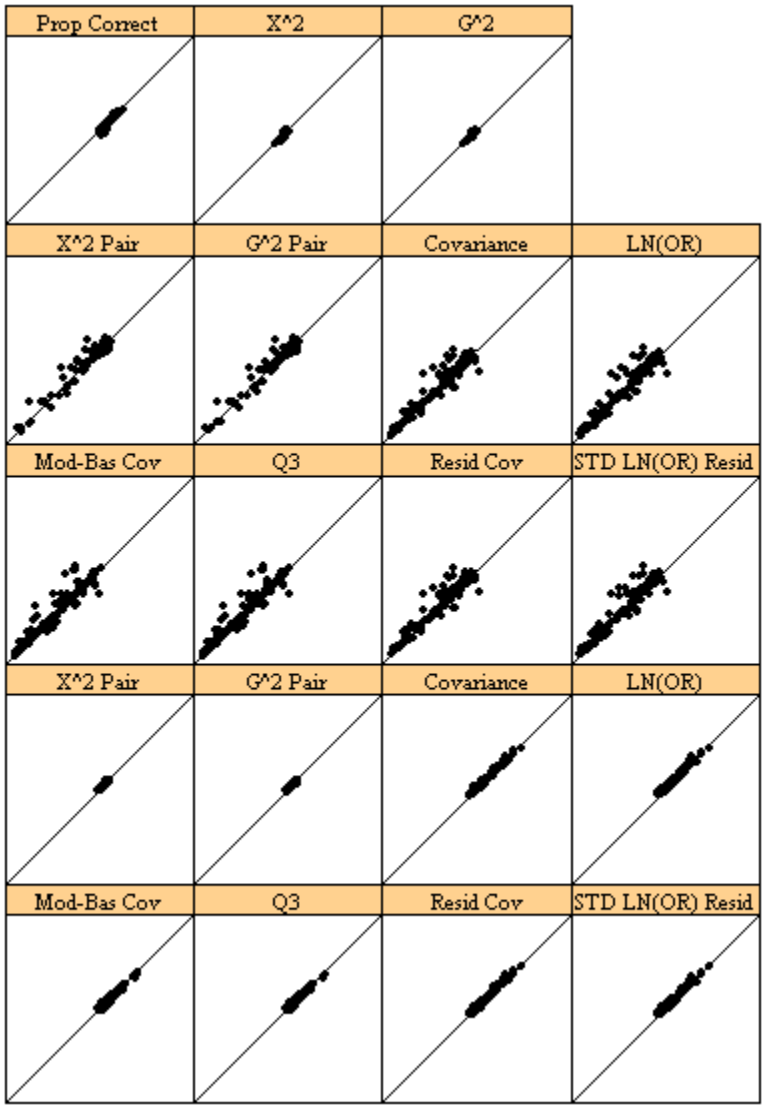
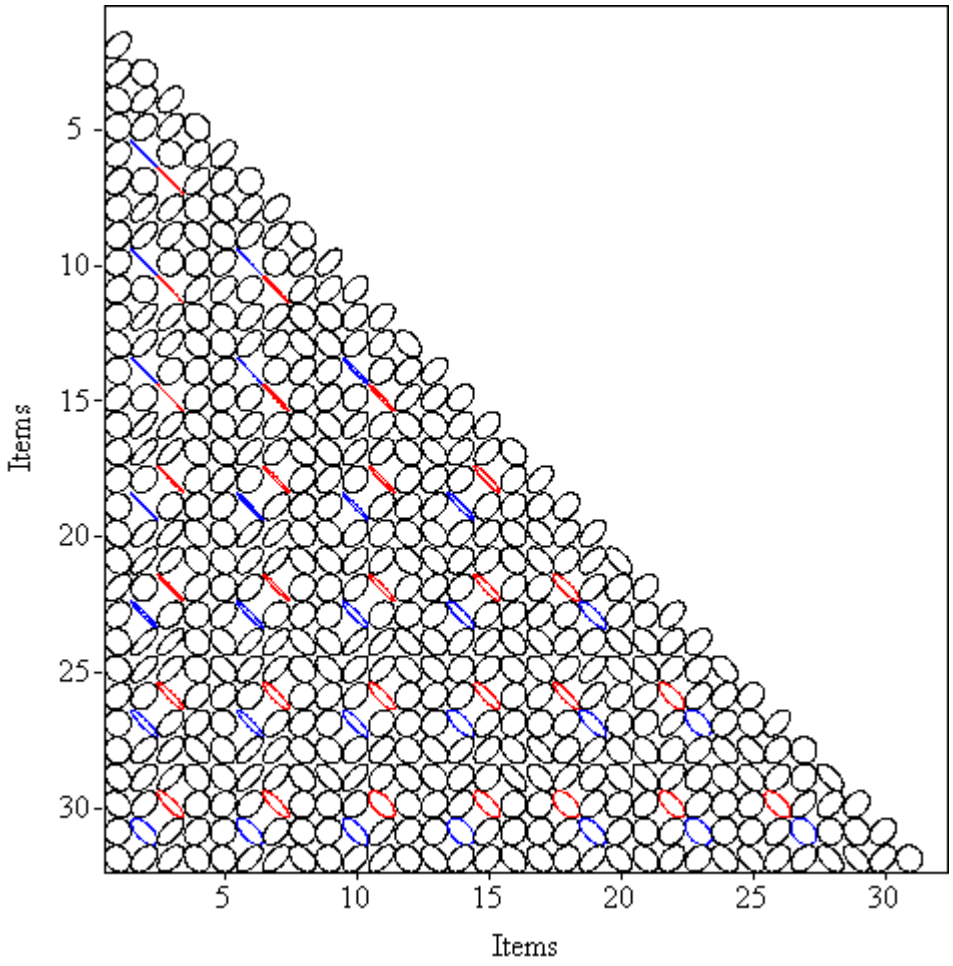


Figure C4: Contour plots of median PPP-values disaggregated by dimension based on conducting PPMC on conjunctive multidimensional data.



However, striking differences are observed by moving around the matrix of contour plots. Consider first the (blue) contour plot for the pairing of items 2 and 6 and the (red) contour plot for the pairing of items 3 and 7. Both are substantially elongated and negatively oriented such that they appear to be lines rather than ellipses. Moving down and to the right throughout the matrix, the contours for multidimensional item-pairs are more and more circular. At the extreme, the contours for the pairing of items 26 and 30 and the pairing of items 27 and 31 are much less sharply oriented. Recall that the

items were ordered from easiest to hardest in terms of their difficulty with respect to the primary dimension (Table 1). This suggests that, for the conjunctive MIRT data, the PPP-values for multidimensional item-pairs depend on the difficulty of the items not only along the auxiliary dimensions (which is explicitly studied), but also upon the difficulty of the items along the *primary* dimension. Further research is necessary to pursue this. Future work may include studies that systematically vary difficulty along each dimension as a manipulated factor.

Returning to the current study, note that an exchangeability assumption regarding θ_2 and θ_3 is still warranted *conditional* on the difficulties of the items that constitute a pair. This was made explicit in Figure 48. In each panel of Figure 48, the median PPP-values do not vary substantially over the auxiliary dimensions after conditioning on the relative values of the item difficulties along the primary dimension.

REFERENCES

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 74,* 255-278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analysis. *Applied Psychological Measurement, 20,* 311-329.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1-23.

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

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of Educational Statistics, 17,* 251-269.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association, 88, 669-679.

Almond, R. G., Dibello, L., Jenkins, F., Mislevy, R. J., Senturk, D., Steinberg, L. S. and Yan, D. (2001). Models for conditional probability tables in educational assessment. In T. Jaakkola and T. Richardson (Eds.), *Artificial intelligence and statistics 2001* (137-143). San Francisco: Morgan Kaufmann.

Almond, R. G., & Mislevy, R. J. (1999). Graphical models and computerized adaptive testing. *Applied Psychological Measurement, 23*, 223–237.

Arminger, G. & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm. *Psychometrika*, *63*, 271-300.

Bartholomew, D. J. (1987). *Latent variable models and factor analysis*. New York: Oxford University Press.

Bayarri, M. J., Berger, J. O. (2000). *P* values for composite null models. *Journal of the American Statistical Association, 95,* 1127–1142.

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

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. W., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing, 4,* 295-301.

Beran, R. & Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Annals of Statistics, 13,* 95-115.

Berkhof, J., van Mechelen, I., & Gelman, A. (2004). Enhancing the performance of a posterior predictive check. IAP Statistics Network Technical Report 0350.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Stine, R. A. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models*. Newbury Park, CA: SAGE Publications.

Bolt, D. M. (2001). Conditional covariance-based representation of multidimensional test structure. *Applied Psychological Measurement, 25,* 244-257.

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.

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

Box, G. E. P. (1980). Sampling and Bayes' inferenence in scientific modelling and robustness. *Journal of the Royal Statistical Society, A, 143,* 383-430.

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

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician, 47,* 69-100.

Brooks, S. P., & Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

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

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

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

Conati, C., Gertner, A. S., VanLehn, K., & Druzdzel, M. J. (1997). On-line student modeling for coached problem solving using Bayesian networks. In *Proceedings of UM-97, Sixth International Conference on User Modeling* (pp. 231–242). Sardinia, Italy: Springer.

Davey, T., Oshima, T. C., & Lee, K. (1996). Linking multidimensional item calibrations. *Applied Psychological Measurement, 20,* 405-416.

Dayton, C. M. (1999). *Latent class scaling analysis*. Quantitative Applications in the Social Sciences Series No. 126. Thousand Oaks, CA: Sage Publications.

de Boeck, P., & Wilson, M. (Eds.) (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

de Finetti, B. (1964). Forsight: Its logical laws, its subjective sources. In H.E. Kyburg & H.E. Smokler (Eds.), *Studies in subjective probability* (pp. 93-158). New York: Wiley.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333-353.

Donoghue, J. R., & Hombo, C. M. (1999). Some asymptotic results on the distribution of an IRT measure of item fit. Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, Kansas, June, 1999.

Donoghue, J. R., & Hombo, C. M. (2001). The effect of item parameter estimation on the distribution of an IRT measure of item fit. Paper presented at the Annual Meeting

of the National Council on Measurement in Education, Seattle, Washington, April, 2001.

Donoghue, J. R., & Hombo, C. M. (2003). A corrected asymptotic distribution of an IRT fit measure that accounts for the effects of item parameter estimation. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, Illinois, April, 2003.

Draper, D. (1996). Comment: Utility, sensitivity analysis, and cross-validation in Bayesian model-checking. *Statistica Sinica, 6,* 760-767.

Edwards, W. (1998). Hailfinder: Tools for and experiences with Bayesian normative modeling. *American Psychologist, 53,* 416-428.

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

Embretson, S. E. (Ed.) (1985). *Test design: Developments in psychology and psychometrics.* Orlando, FL: Academic Press.

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: Springer-Verlag.

Formann, A. K. (1985). Constrained latent class models: Theory and applications. *British Journal of Mathematical and Statistical Psychology, 38,* 87-111.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

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, Montréal, Canada, April, 2005.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 145-161). London: Chapman & Hall.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398-409.

Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice* (pp. 131-143). London: Chapman and Hall.

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

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

Gelman, A. & Rubin, D. B. (1992). Inference from iterative sampling using multiple sequences. *Statistical Science, 7,* 457-511.

Ghosh, M., Ghosh, A., Chen, M.-H., & Agresti, A. (2000). Bayesian estimation for item response models. *Journal of Statistical Planning and Inference, 88,* 99-115.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996a). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 1-19). London: Chapman & Hall.

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

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, B, 29,* 83-100.

Habing, B, & Roussos, L. A. (2003). On the need for local item dependence. *Psychometrika, 68,* 435-451.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Hoijtink, H. (1998). Constrained latent class analysis using the Gibbs sampler and posterior predictive p-values: Applications to educational testing. *Statistica Sinica, 8,* 691-711.

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*. New York: Springer-Verlag.

Hoijtink, H., & Molenaar, I. W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62,* 171-189.

Irvine, S. H., & Kyllonen, P. C. (Eds.) (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Jaynes, E. T. (2003). *Probability theory: The logic of science*. New York: Cambridge University Press.

Jensen, F. V. (1996). *An introduction to Bayesian networks*. New York: Springer-Verlag.

Jensen, F. V. (2001). *Bayesian networks and decision graphs*. New York: Springer-Verlag.

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

Lauritzen, S. L., & Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B, 50,* 157-224.

Lee, S.-Y. (1981). A Bayesian approach to confirmatory factor analysis. *Psychometrika, 46,* 153-160.

Lee, S., & Hershberger, S. L. (1990). A simple rule for generating equivalent models in covariance structure modeling. *Multivariate Behavioral Research, 25*, 313-334.

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research, 39,* 653-686.

Levy, R., & Mislevy, R. J. (2004). Specifying and refining a measurement model for a computer-based interactive assessment. *International Journal of Testing, 4,* 333-369.

Li, Y., Bolt, D. M., & Fu, J. (in press). A comparison of alternative models for testlets. *Applied Psychological Measurement*.

Liebetrau, A. M. (1983). *Measures of association*. Quantitative Applications in the Social Sciences Series No. 32. Thousand Oaks, CA: Sage Publications.

Lindley, D. V., & Novick, M. R. (1981). The role of exchangeability in inference. *The Annals of Statistics, 9,* 45-58.

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1-42.

Lucke, J. F. (2005). The α and ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement, 29,* 65-81.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20,* 389-404.

Maris, G., & Bechger, T. M. (in prepartion). An introduction to the DA-T Gibbs sampler for the two-parameter logistic (2PL) model. *Psicologica*.

Martin, J. D. & VanLehn, K. (1995). A Bayesian approach to cognitive assessment. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141-165). Hillsdale, NJ: Erlbaum.

Martin, J. K., & McDonald, R. P. (1975) Bayesian estimation in unrestricted factor analysis: A treatment for Heywood cases. *Psychometrika, 40,* 505-517.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257-269). New York: Springer-Verlag.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99-114.

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

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300-307.

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

Mislevy, R. J. (1986a). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Mislevy, R. J. (1986b). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics, 11,* 3-31.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59,* 439-483.

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P. Nichols, S. Chipman, & R. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ: Erlbaum.

Mislevy, R. J., Almond, R. G., DiBello, L., Jenkins, F., Steinberg, L. S., Yan, D., & Senturk, D. (2002). Modeling conditional probabilities in complex educational assessments.CSE Technical Report 580. Los Angeles: The National Center for Research on Evaluation, Standards, Student Testing (CRESST), Center for Studies in Education, UCLA.

Mislevy, R.J., & Gitomer, D.H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction, 5,* 253-282.

Mislevy, R. J., & Levy, R. (in press). Bayesian psychometric modeling from an evidence-centered design perspective. In C. R. Rao and S. Sinharay (Eds.) *Handbook of statistics, Volume 17*. North-Holland: Elsevier.

Mislevy, R. J., & Patz, R. J. (1995). *On the consequences of ignoring certain conditional* dependencies in cognitive diagnosis. Paper presented at the Annual Meeting of the American Statistical Association, Orlando, FL, August, 1995.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Newbury Park, CA: Sage.

Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science, 68,* 218-241.

Murdoch, D. J. and Chow, E. D. (1996). A graphical display of large correlation matrices. *The American Statistician, 50,* 178-180.

Muthén, B., & Christofferson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407-419.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18*, 41–68.

Novick, M. R., Jackson, P. H., & Thayer, D. T. (1971). Bayesian inference and the classical test theory model: Reliability and true scores. *Psychometrika, 36,* 261-288.

Patz, R. J., Junker B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146-78.

Patz, R. J., and Junker, B. W. (1999b). Applications and extensions of MCMC for IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Kaufmann.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Logn (Eds.), *Testing structural equation models* (pp. 163-180). Newbury Park, CA: Sage.

Raftery, A. E. (1996). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163-187). London: Chapman & Hall.

Raykov, T., & Penev, S. (1999). On structural equation model equivalence. *Multivariate Behavioral Research*, *34*, 199-244.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M. D. (1997a). 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: Springer-Verlag.

Reckase, M. D. (1997b). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of test items that measure more than one ability. *Applied Psychological Measurement, 15,* 361-373.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 45-57). London: Chapman & Hall.

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.

Roussos, L., & Stout, W. (1996). DIF from the multidimensional perspective. *Applied Psychological Measurement, 20, 335-371*.

Rowe, D. B. (2003). *Multivariate Bayesian statistics: Models for source separation and signal unmixing*. Boca Raton, FL: CRC Press.

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

Rubin, D. B. (1996). Comment: On posterior predictive *p*-values. *Statistica Sinica, 6,* 787-792.

Rubin, D. B., & Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 420-438). Thousand Oaks, CA: Sage.

Rupp, A. A. (2002). Feature selection for choosing and assembling measurement models: A building-block-based organization. *International*

Journal of Testing,

2, 311-360.

Rupp, A. A., Dey, D. K., & Zumbo, B. D. (2004) To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11*, 424-451.

Rupp, A. A., & Mislevy, R. J., (in press). Cognitive foundations of structured item response models. To appear in J.P. Leighton & M. J. Gierl (Eds.), *Cognitive Diagnostic Assessment: Theories and Applications.* Cambridge: Cambridge University Press.

Sahu, S. K. (2002). Bayesian estimation and model choice in item response models. *Journal of Statistical Computation and Simulation, 72,* 217-232.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, 34, (No. 4, Part 2).

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika, 64,* 37-52.

Schum, D. A. (1987). *Evidence and inference for the intelligence analyst.* Lanham, Md.: University Press of America.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: Wiley.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61,* 331-354.

Seltzer, M. H., Wong, W. H., & Bryk, A. S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21,* 131-167.

Shealy, R., & Stout., W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics, 29,* 461-488.

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

Sinharay, S. (in pressa). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*.

Sinharay, S. (in pressb). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics.*

Sinharay, S. Almond, R. G., & Yuan, D. (2004). Assessing fit of models with discrete proficiency variables in educational assessment. ETS research report RR-04-07.

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.

Sinharay, S., & Johnson, M. (2004). Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Sinharay, S., Johnson, M., & Stern, H. S. (in press). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*.

Sinharay, S. & Stern, H. S. (2003). Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference, 111,* 209-221.

Smith, A. F. M., & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, B, 55,* 3-23.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian analysis in expert systems. *Statistical Science, 8,* 219-247.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). *WinBUGS version 1.4: user manual*. Cambridge Medical Research Council Biostatistics Unit. http://www.mrc-bsu.cam.ac.uk/bugs/

Stelzl, I. (1986). Changing a causal hypothesis without changing the fit: Some rules for generating equivalent path models. *Multivariate Behavioral Research*, *21*, 309-331.

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

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52,* 589-617.

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.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7,* 175-192.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50,* 349-364.

Swaminathan, H., & Gifford, J. A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51,* 589-601.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (in press). Assessing the fit of item response theory models. In C. R. Rao and S. Sinharay (Eds.) *Handbook of statistics, Volume 17*. North-Holland: Elsevier.

Sympson, J. B. (1978) A *Model For Testing With Multidimensional Items* En Weiss, D. J. (Ed) Proceedings Of The Computerized Adaptive Testing Conference, Department Of Psychology, University Of Minnesota, Minneapolis.

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

Thurstone, L. L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics, 22,* 1701-1728.

van der Linden, W. J. (1996). Assembling tests for the measurement of multiple traits. *Applied Psychological Measurement, 20,* 373-388.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Vispoel, W. P. (1998). Psychometric characteristics of computer-adaptive and self-adaptive vocabulary tests: The role of answer feedback and test anxiety. *Journal of Educational Measurement, 35,* 155-167.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrika, 50,* 1-26.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Pscychometrika, 45,* 479-494.

Williamson, D. M. (2000). *Utility of model criticism indices for Bayesian inference networks in cognitive assessment*. Unpublished doctoral dissertation, Fordham University, New York.

Williamson, D. M., Bauer, M., Steinberg, L. S., Mislevy, R. J., Behrens, J. T., & DeMark, S. F. (2004). Design rationale for a complex performance assessment. *International Journal of Testing, 4,* 303-332.

Williamson, D. M., Mislevy, R. J., & Almond, R. G. (2001). Model criticism of Bayesian networks with latent variables. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, Washington, April, 2001.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

Yen, W. (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. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64,* 129-152.

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213-249.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 244-256). New York: Springer-Verlag.