CRESST REPORT 840

LIMITED-INFORMATION GOODNESS-OF-FIT TESTING OF DIAGNOSTIC CLASSIFICATION ITEM RESPONSE THEORY MODELS

APRIL 2014

Mark Hansen
Li Cai
Scott Monroe
Zhen Li

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Limited-Information Goodness-of-Fit Testing of Diagnostic Classification Item Response Theory Models

CRESST Report 840

Mark Hansen, Li Cai, Scott Monroe, and Zhen Li
CRESST/University of California, Los Angeles

May 2014

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Mark Hansen is supported by a dissertation improvement grant from the National Science Foundation (#1260746). Li Cai is supported by grants from the Institute of Education Sciences (R305B080016) and National Institute on Drug Abuse (R01DA026943 and R01DA030466). The views expressed here belong to the authors and do not reflect the views or policies of the funding agencies or grantees.

Address all correspondence to: Li Cai, CRESST, GSE&IS, UCLA, Los Angeles, CA, USA 90095-1521. Email: lcai@ucla.edu. Phone: 310.206.0583 Fax: 310.206.5830

To cite from this report, please use the following as your APA reference: Hansen, M., Monroe, S., & Cai, L., & Li, Z. (2014). *Limited-information goodness-of-fit testing of diagnostic classification item response theory models.* (CRESST Report 840). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST.

TABLE OF CONTENTS

LIMITED-INFORMATION GOODNESS-OF-FIT TESTING OF DIAGNOSTIC CLASSIFICATION ITEM RESPONSE THEORY MODELS

Mark Hansen, Li Cai, Scott Monroe, and Zhen Li
CRESST/University of California, Los Angeles

Abstract

It is a well-known problem in testing the fit of models to multinomial data that the full underlying contingency table will inevitably be sparse for tests of reasonable length and for realistic sample sizes. Under such conditions, full-information test statistics such as Pearson's X^2 and the likelihood ratio statistic G^2 are poorly calibrated. Limited-information fit statistics, such as the M_2 statistic proposed by Maydeu-Olivares & Joe (2006), have been suggested as possible alternatives to full-information tests in various modeling including item response theory models. In this study, we considered the application of M_2 to the goodness-of-fit testing of diagnostic classification models (e.g., Rupp, Templin, & Henson, 2010). Through a series of simulation studies, we found that M_2 is well calibrated across a range of diagnostic model structures. The sensitivity of the test statistic to detect various types of model misspecification was also examined. M_2 was found to be sensitive to certain misspecifications of the item model (e.g., fitting disjunctive models to data generated according to a conjunctive model), errors in the Q-matrix, and local item dependence due to unmodeled testlet effects. On the other hand, M_2 was largely insensitive to misspecifications in the distribution of higher-order latent dimensions and to the specification of an extraneous attribute. To complement the analyses of overall model goodness-of-fit, we investigated the potential utility of the Chen and Thissen (1997) local dependence statistic X_{LD}^2 for characterizing sources of misfit. The X_{LD}^2 statistic was found to be slightly conservative (with Type I error rates consistently below the nominal level) but still useful in identifying potential misspecifications. Patterns of local dependence arising due to model misspecifications are illustrated. Finally, we used the M_2 and X_{LD}^2 statistics to evaluate a diagnostic model fit to a data from the Trends in Mathematics and Science Study (TIMSS), drawing upon analyses previously conducted by Lee, Park, and Taylan (2011).

1 Introduction

Diagnostic classification models (see, e.g., Rupp, Templin, & Henson, 2010) have received increasing attention within the field of educational and psychological measurement. The popularity of these models may be largely due to their perceived ability to provide useful information concerning both examinees (classifying them according to their attribute profiles) and test items (describing the particular attributes that are relevant to or required in order to achieve a certain response). Despite these attractive features, it is important to note the potential for biased interpretations when such models are misspecified. Various authors have noted that methods for evaluating the fit of diagnostic models remain relatively underdeveloped (e.g., Maris & Bechger, 2009; Wilhelm & Robitzsch, 2009; Rupp et al., 2010). That said, there has been notable progress (e.g., Sinharay & Almond, 2009; Lai, Cui, & Gierl, 2012; de la Torre, 2008; Rupp & Templin, 2008; Kunina-Habenicht, Rupp, & Wilhelm, 2012).

In this study, we examine the utility of limited information goodness-of-fit statistics (e.g., Bartholomew & Leung, 2002) for evaluating the fit of diagnostic classification models. Specifically, we apply the M_2 statistic proposed by Maydeu-Olivares & Joe (2006) to a range of diagnostic model structures and consider its calibration and sensitivity across a range of misspecifications. Various M_2-type statistics have been applied to an increasing assortment of item response theory models (e.g., Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Joe & Maydeu-Olivares, 2010; Cai & Hansen, 2013; Cai & Monroe, 2014). Limited-information fit statistics have been suggested as a possible approach for evaluating the use of diagnostic models (e.g., Rupp et al., 2010), and initial applications appear quite promising (Templin, 2007; Jurich, Bradshaw, & DeMars, 2014). Here, we seek to continue to develop this line of work. We begin by describing the development of the statistic within the context of diagnostic modeling, then evaluate its performance through a series of simulation studies. We then use the statistic to assess the fit of a diagnostic model to real data. Finally, we discuss some of the limitations and opportunities to further develop this work.

2 General Diagnostic Classification Modeling Framework

2.1 An Item Response Model For Diagnostic Classification

In this section, we describe a higher-order, hierarchical diagnostic classification model (Cai, 2013a) to which we will apply the limited-information statistic. This model may be best understood as an extension of existing diagnostic modeling frameworks, such as the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009), which we use here as a starting point. Let there be a total of $i = 1, \ldots, I$ items. In this research we limit the scope to dichotomously scored items, noting that extending the theory to multiple-categorical

data is straightforward. Let the response categories be coded as $c = 0$ (incorrect/no endorsement) or $c = 1$ (correct/endorsement). Let there be K dichotomous (0—1) underlying latent attribute variables, $\boldsymbol{x} = (x_1, \dots, x_k, \dots, x_K)'$, where $\mathrm{x_k} = 1$ indicates mastery/possession of an attribute. The relationship between items and attributes is captured by the Q-matrix, which is $I \times K$ matrix of zeros and ones. The $(\mathrm{i}, \mathrm{k})^{\mathrm{th}}$ entry in the Q-matrix is denoted as $\mathrm{q_{ik}}$, and takes on a value of one if item i measures attribute k.

Let $T_i(1|\boldsymbol{x})$ be the category 1 response function for item i:

$$T_i(c|\boldsymbol{x}) = \frac{1}{1 + \exp(-\eta_i)}, \tag{1}$$

where the linear predictor is

$$\eta_i = \alpha_i + \boldsymbol{\lambda}_i' h_i(\mathbf{Q}, \boldsymbol{x}). \tag{2}$$

It follows that for category 0, the response function is

$$T_i(0|\boldsymbol{x}) = 1.0 - T_i(1|\boldsymbol{x}). \tag{3}$$

The α's and λ's in the linear predictor are the item parameters, and $h_i(\mathbf{Q}, \boldsymbol{x})$ is a potentially vector-valued function that defines how the measured attributes combine to create the linear predictor portion of the item response model in Equation (2). As noted by Rupp et al. (2010) and Choi, Rupp, & Pan (2013), placing certain constraints on the λ's yields several of the more commonly utilized diagnostic models. For instance, suppose that according to the Q-matrix item i measures attributes 1 and 2, and that successful solution of the item requires both attributes. The linear predictor may take the following deterministic-input noisy "and" gate (DINA; Junker & Sijtsma, 2001) form with a single free parameter for the interaction term:

$$\boldsymbol{\eta_i} = \boldsymbol{\alpha_i} + \mathbf{0}\boldsymbol{x_1} + \mathbf{0}\boldsymbol{x_2} + \boldsymbol{\lambda_i x_1 x_2}, \tag{4}$$

in which case $h_i(\mathbf{Q}, \boldsymbol{x}) = (x_1, x_2, x_1 x_2)'$ contains the main effects and the second-order interaction, and $\boldsymbol{\lambda}_i = (0, 0, \lambda_i)'$ contains fixed parameters. Or perhaps the item only requires the mastery of either attribute 1 or attribute 2. Then the linear predictor of the IRT model may take the following deterministic-input noisy "or" gate (DINO; Templin & Henson, 2006) form with a single slope parameter set equal between the main effects and the interaction term in absolute magnitude, albeit the interaction term is of the opposite direction:

$$\eta_i = \alpha_i + \lambda_i x_1 + \lambda_i x_2 - \lambda_i x_1 x_2, \tag{5}$$

in order to reflect the specification that mastery of both attributes does not lead to a further increase in the logit, and in this case $\boldsymbol{\lambda}_i = (\lambda_i, \lambda_i, -\lambda_i)'$ contains linear restrictions.

Yet another possibility is that each attribute contributes to some increase in the logit, and that the magnitude of the increase due to one attribute does not depend on the mastery of the other. In that case, the linear predictor might contain only the main effect terms, and the model would take the form of von Davier's (2005) general diagnostic model or the compensatory reparameterized unified model (C-RUM; Hartz, 2002):

$$\eta_i = \alpha_i + \lambda_i x_1 + \lambda_i x_2. \tag{6}$$

Let Y_i be a random variable whose realization y_i is a response to item i. Regardless of the exact form of the model or the number of attributes, the probability mass function of Y_i, conditional on $\boldsymbol{x}$, is that of a Bernoulli:

$$P(Y_i = y_i|\boldsymbol{x}) = [T_i(y_i|\boldsymbol{x})]^{y_i}[1.0 - T_i(y_i|\boldsymbol{x})]^{1-y_i}. \tag{7}$$

2.2 Hierarchical and Higher-Order Extensions

Conditional independence is a critical assumption for model building in all of IRT analysis. In the case of DCMs, it is customary to assume the independence of item responses given the attributes (e.g., Templin & Henson, 2006). That is, the conditional response pattern probability factors:

$$\pi(\boldsymbol{y}|\boldsymbol{x}) = P\left(\bigcap_{i=1}^{I} Y_i = y_i \,\middle|\, \boldsymbol{x}\right) = \prod_{i=1}^{I} P(Y_i = y_i|\boldsymbol{x}), \tag{8}$$

where $\boldsymbol{y} = (y_1, \dots, y_I)'$ is an $I \times 1$ vector that contains the observed response pattern. However, this is unrealistic if item i happens to belong to a cluster of items dependent on the same stimulus in a passage-based reading assessment, or if it falls into a specific content subdomain, e.g., social aspects of quality of life, along with other items. A standard strategy in item factor analysis is to include additional random effects to account for potential residual dependence due to the common source of variation shared by a set of items. Let there be $s = 1, \dots, S$ such clusters of items. Furthermore, if we assume that the clusters are mutually exclusive, the response function for item i in category 1 becomes

$$T_i(1|\boldsymbol{x}, \xi_s) = \frac{1}{1 + \exp[-(\alpha_{ic} + \boldsymbol{\lambda}_i' h_i(\mathbf{Q}, \boldsymbol{x}) + \beta_s \xi_s)]}, \tag{9}$$

where β_s is the item slope on specific factor/dimension ξ_s. Again the response function for category 0 becomes

$$T_i(0|\boldsymbol{x}, \xi_s) = 1.0 - T_i(1|\boldsymbol{x}, \xi_s). \tag{10}$$

The model resembles a bifactor model or testlet model (Gibbons & Hedeker, 1992). An item is permitted to load on at most one specific dimension. In this case, conditional

independence may be more amenable:

$$\pi(\boldsymbol{y}|\boldsymbol{x},\xi_1,\dots,\xi_S) = P\left(\bigcap_{i=1}^{I} Y_i = y_i \,\middle|\, \boldsymbol{x},\xi_1,\dots,\xi_S\right) = \prod_{s=1}^{S}\prod_{i\in\mathfrak{H}_s} P(Y_i = y_i|\boldsymbol{x},\xi_s), \tag{11}$$

where $\mathfrak{H}_s$ is a notational shorthand for the set of items that load on specific dimension s, and $P(Y_i = y_i|\boldsymbol{x},\xi_s) = [T_i(y_i|\boldsymbol{x},\xi_s)]^{y_i}[T_i(y_i|\boldsymbol{x},\xi_s)]^{1-y_i}$ is again a Bernoulli probability mass function.

Suppose the distribution of the ξ's are given by $g(\xi_1|\boldsymbol{x})g(\xi_2|\boldsymbol{x})\cdots g(\xi_S|\boldsymbol{x})$, i.e., the specific dimensions are conditionally independent given $\boldsymbol{x}$. We may integrate all S specific dimensions out without a full S-dimensional integral. This is because we may utilize the familiar dimension reduction method (see Cai, Yang, & Hansen, 2011; Rijmen, 2009) developed for item bifactor analysis to transform the following S-fold integral

$$\pi(\boldsymbol{y}|\boldsymbol{x}) = \int \pi(\boldsymbol{y}|\boldsymbol{x},\xi_1,\dots,\xi_S)\, g(\xi_1|\boldsymbol{x})\cdots g(\xi_S|\boldsymbol{x})d\xi_1\cdots d\xi_S, \tag{12}$$

into a series of one-dimensional integrations

$$\begin{aligned}\pi(\boldsymbol{y}|\boldsymbol{x}) &= \int \prod_{s=1}^{S}\prod_{i\in\mathfrak{H}_s} P(Y_i = y_i|\boldsymbol{x},\xi_s)\, g(\xi_1|\boldsymbol{x})\cdots g(\xi_S|\boldsymbol{x})d\xi_1\cdots d\xi_S \\ &= \prod_{s=1}^{S}\int \prod_{i\in\mathfrak{H}_s} P(Y_i = y_i|\boldsymbol{x},\xi_s)\, g(\xi_s|\boldsymbol{x})d\xi_s,\end{aligned} \tag{13}$$

which will vastly reduce the amount of time needed for maximum marginal likelihood based parameter estimation because these integrals must be numerically evaluated.

As per de la Torre & Douglas (2004), we may further model the latent attribute profiles for individuals by regressing the x's on p higher-order dimensions $\boldsymbol{\theta}$. For example, we may use a multidimensional extension of the 2-parameter logistic model (Reckase, 2009) to relate the latent attributes to the latent dimensions:

$$P(x_k = 1|\boldsymbol{\theta}) = \pi_k(\boldsymbol{\theta}) = \frac{1}{1+\exp\left[-(a_{k0}+a_{k1}\theta_1+\cdots+a_{kp}\theta_p)\right]}. \tag{14}$$

Again, if we assume conditional independence of the latent attributes given $\boldsymbol{\theta}$, we may write

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K}[\pi_k(\boldsymbol{\theta})]^{x_k}[1-\pi_k(\boldsymbol{\theta})]^{1-x_k}. \tag{15}$$

When we combine $\pi(\boldsymbol{y}|\boldsymbol{x})$ from Equation (11) with $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ from Equation (14), we see that the contribution to marginal likelihood from response pattern $\boldsymbol{y}$ can be obtained as:

$$\pi(\boldsymbol{y}) = \int \left[\int \pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{x} \right] g(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{16}$$

where the integral in the brackets is actually a 2^K-term summation over the (conditional) attribute profile probabilities for all $\boldsymbol{x}$.

3 Limited-information Goodness-of-fit Testing

3.1 Maximum Marginal Likelihood Estimation

Let $\boldsymbol{\gamma}$ be a $d\times1$ vector that collects together all free parameters in the model. These include parameters from all I items (the α's and λ's), parameters for the distribution of the specific dimensions $g(\xi_s|\boldsymbol{x})$, the higher-order IRT model (the a's), and the parameters from the distribution of the higher-order dimensions $g(\boldsymbol{\theta})$. To emphasize the fact that the marginal likelihood in Equation (16) is a function of the parameters once the response pattern is observed (and considered fixed), let $\pi_{\boldsymbol{y}}(\boldsymbol{\gamma})$ denote the marginal likelihood.

For I items, the IRT model generates a total of $C = 2^I$ cross-classifications or possible item response patterns in the form of a contingency table. Based on a sample of N respondents, let the observed proportion associated with pattern $\boldsymbol{y}$ be denoted as $p_{\boldsymbol{y}}$. The sampling model for this contingency table is a multinomial distribution with C cells and N trials. The multinomial log-likelihood for the item parameters $\boldsymbol{\gamma}$ is proportional to

$$\log L(\boldsymbol{\gamma}) \propto N \sum_{\boldsymbol{y}} p_{\boldsymbol{y}} \log \pi_{\boldsymbol{y}}(\boldsymbol{\gamma}), \tag{17}$$

where the summation is over all C response patterns. Maximization of the log-likelihood (e.g., with the EM algorithm; Bock & Aitkin, 1981) leads to the maximum marginal likelihood estimator $\widehat{\boldsymbol{\gamma}}$.

Upon finding $\widehat{\boldsymbol{\gamma}}$, the model generates model-implied probabilities for each response pattern $\hat{\pi}_{\boldsymbol{y}} = \pi_{\boldsymbol{y}}(\widehat{\boldsymbol{\gamma}})$. Suppose the model-implied response pattern probabilities $\hat{\pi}_{\boldsymbol{y}}$ are collected into a $C\times1$ vector $\widehat{\boldsymbol{\pi}}$ of all model-implied response pattern probabilities. By analogy, let a $C\times1$ vector $\boldsymbol{\pi}_*$ contain the true (population) response pattern probabilities. Similarly, the observed proportions $p_{\boldsymbol{u}}$ can be collected into a $C\times1$ vector $\boldsymbol{p}$. For example, for 3 items there are $2^3 = 8$ item response patterns, and the response pattern probabilities

and observed proportions are:

$$\boldsymbol{\pi}_* = \begin{pmatrix} \pi_{000} \\ \pi_{001} \\ \pi_{010} \\ \pi_{011} \\ \pi_{100} \\ \pi_{101} \\ \pi_{110} \\ \pi_{111} \end{pmatrix}, \qquad \widehat{\boldsymbol{\pi}} = \begin{pmatrix} \hat{\pi}_{000} \\ \hat{\pi}_{001} \\ \hat{\pi}_{010} \\ \hat{\pi}_{011} \\ \hat{\pi}_{100} \\ \hat{\pi}_{101} \\ \hat{\pi}_{110} \\ \hat{\pi}_{111} \end{pmatrix} = \begin{pmatrix} \pi_{000}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{001}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{010}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{011}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{100}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{101}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{110}(\widehat{\boldsymbol{\gamma}}) \\ \pi_{111}(\widehat{\boldsymbol{\gamma}}) \end{pmatrix}, \qquad \boldsymbol{p} = \begin{pmatrix} p_{000} \\ p_{001} \\ p_{010} \\ p_{011} \\ p_{100} \\ p_{101} \\ p_{110} \\ p_{111} \end{pmatrix}. \tag{18}$$

Using this setup, exactly correct model specification (i.e., the model fits perfectly) in the population can be understood as the statement that there exists $\boldsymbol{\gamma}_*$ such that $\boldsymbol{\pi}(\boldsymbol{\gamma}_*) = \boldsymbol{\pi}_*$. The values $\boldsymbol{\gamma}_*$ may be taken as the true parameters to be estimated.

Under correct model specification, from results in discrete multivariate analysis (e.g., Bishop, Fienberg, & Holland, 1975), the maximum likelihood estimator is consistent, asymptotically normal and asymptotically efficient, which can be summarized as follows:

$$\sqrt{N}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_*) \xrightarrow{D} \mathcal{N}_d(\mathbf{0}, \mathcal{F}^{-1}), \tag{19}$$

where $\mathcal{F} = \boldsymbol{\Delta}_*' \, [diag(\boldsymbol{\pi}_*)]^{-1} \boldsymbol{\Delta}_*$ is the $d \times d$ Fisher information matrix, with the Jacobian matrix $\boldsymbol{\Delta}_*$ defined as the $C \times d$ matrix of all first-order partial derivatives of the response pattern probabilities with respect to the parameters, evaluated at $\boldsymbol{\gamma}_*$:

$$\boldsymbol{\Delta}_* = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial \boldsymbol{\gamma}'}.$$

3.2 Distribution of Residuals under Maximum Likelihood Estimation

It can be shown that the asymptotic distribution of $(\boldsymbol{p} - \boldsymbol{\pi}_*)$ is C-variate normal:

$$\sqrt{N}(\boldsymbol{p} - \boldsymbol{\pi}_*) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Xi}), \tag{20}$$

where $\boldsymbol{\Xi} = diag(\boldsymbol{\pi}_*) - \boldsymbol{\pi}_* \boldsymbol{\pi}_*'$ is the multinomial covariance matrix. The cell residual vector $(\mathrm{p}\text{-}\hat{\pi})$ is asymptotically C-variate normal under maximum likelihood estimation:

$$\sqrt{N}(\boldsymbol{p} - \widehat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_C(\mathbf{0}, \boldsymbol{\Gamma}), \tag{21}$$

where $\boldsymbol{\Gamma} = \boldsymbol{\Xi} - \boldsymbol{\Delta}_* \mathcal{F}^{-1} \boldsymbol{\Delta}_*'$.

The model also generates implied marginal probabilities. Consider the 3-item example from above. There are 3 first order marginal probabilities $\dot{\pi}_i$ $(i = 1,2,3)$, one per item. There are also 3 second order marginal probabilities $\ddot{\pi}_{ii'}$ for the unique item pairs ($1 \leq i' < i \leq 3$). In general, these probabilities correspond to the I sets of univariate and $I(I-1)/2$ sets

of bivariate margins that can be obtained from the full C-dimensional contingency table using a reduction operator matrix (see e.g., Maydeu-Olivares & Joe, 2006). An example is given below:

$$\widehat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \ddot{\pi}_{21} \\ \ddot{\pi}_{31} \\ \ddot{\pi}_{32} \end{pmatrix} = \mathbf{L}\widehat{\boldsymbol{\pi}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \widehat{\pi}_{000} \\ \widehat{\pi}_{001} \\ \widehat{\pi}_{010} \\ \widehat{\pi}_{011} \\ \widehat{\pi}_{100} \\ \widehat{\pi}_{101} \\ \widehat{\pi}_{110} \\ \widehat{\pi}_{111} \end{pmatrix}, \tag{22}$$

where $\mathbf{L}$ is a fixed operator matrix of 0s and 1s that reduces the response pattern probabilities and proportions into marginal probabilities and proportions up to order 2. The vector $\widehat{\boldsymbol{\pi}}_2 = \mathbf{L}\widehat{\boldsymbol{\pi}} = \mathbf{L}\boldsymbol{\pi}(\widehat{\boldsymbol{\gamma}}) = \boldsymbol{\pi}_2(\widehat{\boldsymbol{\gamma}})$ contains all first and second order model-implied marginal probabilities, evaluated at the maximum likelihood estimate. Obviously $\boldsymbol{p}_2 = \mathbf{L}\boldsymbol{p}$ is the vector of first and second order observed marginal proportions.

A requirement on $\mathbf{L}$ is that it has full row rank. It implies that the marginal residual vector $(\mathrm{p}_2\text{-}\widehat{\pi}_2) = \mathrm{L}(\mathrm{p}\text{-}\widehat{\pi})$ is a full rank linear transformation of the multinomial cell residual vector $(\boldsymbol{p} - \widehat{\boldsymbol{\pi}})$. Therefore, the marginal residual vector $(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2)$ is also asymptotically normal:

$$\sqrt{N}(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2) = \sqrt{N}\mathbf{L}(\boldsymbol{p} - \widehat{\boldsymbol{\pi}}) \xrightarrow{D} \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Gamma}_2), \tag{23}$$

and $\boldsymbol{\Gamma}_2 = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}' = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}' - \mathbf{L}\boldsymbol{\Delta}_*\mathcal{F}^{-1}\boldsymbol{\Delta}_*'\mathbf{L}' = \boldsymbol{\Xi}_2 - \boldsymbol{\Delta}_{2*}\mathcal{F}^{-1}\boldsymbol{\Delta}_{2*}'$, where $\boldsymbol{\Xi}_2 = \mathbf{L}\boldsymbol{\Xi}\mathbf{L}'$, and $\boldsymbol{\Delta}_{2*} = \mathbf{L}\boldsymbol{\Delta}_*$ is the Jacobian of the marginal probabilities:

$$\boldsymbol{\Delta}_{2*} = \mathbf{L}\frac{\partial\boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}'} = \frac{\partial\mathbf{L}\boldsymbol{\pi}(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}'} = \frac{\partial\boldsymbol{\pi}_2(\boldsymbol{\gamma}_*)}{\partial\boldsymbol{\gamma}'}.$$

The dimensionality r of the multivariate normal random vector is equal to the rank of $\mathbf{L}$. The rank of $\boldsymbol{\Delta}_{2*}$ determines local identification of the model from the marginal probabilities. If $\boldsymbol{\Delta}_{2*}$ has full column rank, the model is locally identified. For dichotomously scored item responses, $r = I + I(I-1)/2$.

3.3 A Proposed Test Statistic

The full-information test statistics such as likelihood ratio G^2 and Pearson's X^2 use residuals based on the full response pattern cross-classifications to test the fit of the model against the general multinomial alternative. The comparison between $\widehat{\pi}_{\mathrm{u}}$ and $p_{\boldsymbol{u}}$ (on logarithmic or linear scales) leads to well-known goodness of fit statistics such as the likelihood ratio G^2 and Pearson's X^2:

$$G^2 = 2N \sum_{\boldsymbol{u}} p_{\boldsymbol{u}} \log \frac{p_{\boldsymbol{u}}}{\hat{\pi}_{\boldsymbol{u}}}, \qquad X^2 = N \sum_{\boldsymbol{u}} \frac{(p_{\boldsymbol{u}} - \hat{\pi}_{\boldsymbol{u}})^2}{\hat{\pi}_{\boldsymbol{u}}}. \tag{24}$$

Under the null hypothesis that the IRT model fits exactly, these two statistics have the same asymptotic reference distribution, which is a central chi-square with degrees-of-freedom equal to C-1-d (Bishop et al., 1975).

Unfortunately as the number of items increases, the number of response patterns increases exponentially. For more than a dozen or so dichotomous items, the contingency table upon which the multinomial is defined becomes sparse for most realistic *N*. It is well known that the asymptotic chi-square approximations for the full-information test statistics break down under sparseness (see e.g., Bartholomew & Tzamourani, 1999), and the utility of the full-information overall goodness of fit indices for routine IRT applications is questionable at best.

Recently, limited-information overall fit statistics such as Maydeu-Olivares and Joe's (2006) M_2 have been developed. Limited-information fit statistics use residuals based on lower order (e.g., first and second order) margins of the contingency table. These lower order margins are far better filled when compared to the sparse full contingency table. There is growing awareness that limited-information tests can maintain correct size and can be more powerful than full-information tests (Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Joe & Maydeu-Olivares, 2010; Cai & Hansen, 2013). Moreover, Templin (2007) and Jurich, Bradshaw, & DeMars (2014) have demonstrated the potential usefulness of limited-information statistics in evaluating the fit of diagnostic classification models, in particular. We are not aware, however, of a comprehensive study of the theoretical and empirical aspects of the limited-information fit testing approach for diagnostic classification models, over a wide range of model misspecifications, under reasonably realistic conditions, and implementing both the overall test and diagnostic indices in a publicly available software distribution. Let $\widehat{\boldsymbol{\Xi}} = diag(\widehat{\boldsymbol{\pi}}) - \widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{\pi}}'$ denote the multinomial covariance matrix evaluated at $\widehat{\boldsymbol{\gamma}}$, and let $\widehat{\boldsymbol{\Xi}}_2 = \mathbf{L}\widehat{\boldsymbol{\Xi}}\mathbf{L}'$. Also evaluate the marginal Jacobian at $\widehat{\boldsymbol{\gamma}}$,

$$\widehat{\boldsymbol{\Delta}}_2 = \frac{\partial \boldsymbol{\pi}_2(\widehat{\boldsymbol{\gamma}})}{\partial \boldsymbol{\gamma}'}.$$

When $\widehat{\boldsymbol{\Delta}}_2$ has full column rank, the statistic

$$M_2 = N(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2)'\widehat{\boldsymbol{\Omega}}\,(\boldsymbol{p}_2 - \widehat{\boldsymbol{\pi}}_2), \tag{25}$$

where $\widehat{\Omega} = \widehat{\Xi}_2^{-1} - \widehat{\Xi}_2^{-1}\widehat{\Delta}_2\left(\widehat{\Delta}_2'\widehat{\Xi}_2^{-1}\widehat{\Delta}_2\right)^{-1}\widehat{\Delta}_2'\widehat{\Xi}_2^{-1}$, is asymptotically chi-square distributed with r-d degrees-of-freedom under the null hypothesis that the model fits exactly in the population

(Browne, 1984). To see this is the case, we first recognize that from Equation (14), $\sqrt{N}(p_2\text{-}\hat{\pi}_2)$ is asymptotically a normal random vector with zero means and covariance matrix $\Xi_2\text{-}\Delta_{2*}F^{-1}\Delta_{2*}^{'}$. By the continuous mapping theorem and the consistency of the maximum likelihood estimator, $\hat{\Omega}$ converges in probability to the limiting weight matrix $\lim_{N\to\infty}\hat{\Omega} = \Omega$, where $\Omega = \Xi_2^{-1}\text{-}\Xi_2^{-1}\Delta_{2*}\left(\Delta_{2*}^{'}\Xi_2^{-1}\Delta_{2*}\right)^{-1}\Delta_{2*}^{'}\Xi_2^{-1}$. The product $\left(\Xi_2\text{-}\Delta_{2*}F^{-1}\Delta_{2*}^{'}\right)\Omega = I_r\text{-}\Xi_2^{-1}\Delta_{2*}\left(\Delta_{2*}^{'}\Xi_2^{-1}\Delta_{2*}\right)^{-1}\Delta_{2*}^{'}$ is idempotent and its rank is equal to r-d. Therefore, by Cochran's theorem and Slutsky's theorem, M_2 is asymptotically chi-squared with r-d degrees-of-freedom.

When the model does not fit exactly in the population, the quadratic form in M_2 provides a mechanism for computing a Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993; Maydeu-Olivares, 2013) type index to characterize the degree of model error. This is because the limiting mean of $\sqrt{N}\left(p_2\text{-}\hat{\pi}_2\right)$ will no longer be zero when there does not exist a γ_* such that $\pi(\gamma_*) = \pi_*$. Let $\hat{F} = \left(p_2\text{-}\hat{\pi}_2\right)^{'}\hat{\Omega}\left(p_2\text{-}\hat{\pi}_2\right)$ be the observed noncentrality. As per Browne and Cudeck (1993), an unbiased estimate of the population noncentrality is $F_* = \hat{F}\text{-}df/N$. The sample RMSEA based on M_2 is defined as a measure of the per degree-of-freedom noncentrality (truncated at 0):

$$\text{RMSEA} = \max\left(\sqrt{\frac{F_*}{df}}, 0\right). \tag{26}$$

Confidence intervals of RMSEA may be easily computed from the noncentral chi-square distribution by following established procedures in Browne and Cudeck (1993).

4 Simulation Studies

In order to evaluate the performance of M_2 in the context of diagnostic classification modeling, we conducted a series of simulation studies. First, we evaluated the calibration of the test statistic when the fitted model was correctly specified (i.e., matched the generating model; the null condition). We then examined the power of M_2 to detect a variety of model misspecifications.

For all simulation conditions, the test length was 24 items, and there were four latent attributes. A Q-matrix was obtained by randomly assigning each item to load onto exactly two of the four attributes. This random assignment was balanced, such that all two-attribute combinations were represented equally within the test. A higher-order DINA model was used in all data generation. Item parameters were randomly drawn from distributions of specified

values that are typical of those found in empirical analyses of educational assessment data.

Rather than sampling slopes and intercepts directly, distributions were instead obtained by specifying distributions of correct response probabilities for respondents that either lack or possess the requisite underlying attributes (i.e., "guessing" and "slipping," respectively). These values were then transformed to match the LCDM parameterization (Henson et al., 2009). Guessing parameters were drawn from a beta distribution with a mean of 0.2 and standard deviation of 0.05. The sampling distribution for the slipping parameters had a mean of 0.10 and standard deviation of 0.05. Item intercepts (α_i) were computed from the guessing parameters (g_i) in the following manner:

$$\alpha_i = -\log\left(\frac{1}{g_i - 1}\right)$$

This intercept, together with the slipping parameter (s_i), was then used to obtain the slope parameter:

$$\gamma_i = -\log\left(\frac{1}{1 - s_i} - 1\right) - \alpha_i$$

In addition to the latent attribute variables, items were also influenced in some data generating conditions by six group-specific dimensions (i.e., testlet effects). The slopes of the items on these dimensions were equal within a data generating condition, $\beta = (0, 1, 2)$.

Various higher-order structures were used to generate the distribution of the latent attributes (the x variables). Models with a single higher-order dimension utilized a one-parameter logistic (1PL) IRT model, with slope $a = 1.5$ for all items and intercepts of $c_1 = -0.45, c_2 = 0.15, c_3 = -0.15, c_4 = 0.45$. Scores for the higher-order dimension were sampled from either a standard normal density or from one of four non-normal densities illustrated in Figure 1. The non-normal distributions were parameterized as Davidian curves (Woods and Lin, 2009), each with mean 0 and variance 1. The four densities can be described as Bimodal ($\partial_1 = -0.10, \partial_2 = 1.98$), Extreme Bimodal ($\partial_1 = -0.52, \partial_2 = 2.29$), Right-skewed ($\partial_1 = 0.69, \partial_2 = 4.09$), and Extreme Right-skewed ($\partial_1 = .79, \partial_2 = 6.58$), where ∂_1 and ∂_2 are the skewness and kurtosis coefficients, respectively. These densities are similar to those used previously in research on non-normal latent trait density estimation in IRT (see, e.g., Woods & Lin, 2009).

Insert Figure 1 about here

In addition to the univariate normal and non-normal higher-order distributions, we also generated data from a model in which the higher-order structure was multidimensional. For these conditions, attributes 1 and 2 loaded onto one dimension (θ_1), and attributes 3 and 4 loaded onto a second dimension (θ_2). The means of the two higher-order dimensions were 0,

their variances were 1, and the correlation between domains varied across conditions ($\rho = 0.4, 0.6, 0.8$).

The Q-matrix and the item parameters used in data generation are presented in Table 1. Figure 2 presents path diagrams for the three basic model structures: a higher-order DINA model (top panel), a DINA model with correlated higher-order dimensions (middle panel), and a higher-order DINA with testlet effects (bottom panel). For each data generating condition, 500 datasets were generated in three sample sizes ($\mathrm{N} = 500, 1000, 2000$). The fitted models used with each data generating condition are summarized in Table 2. The first rows of Table 2 describe the null conditions, and the remaining sections identify the various misspecified models fit within each generating condition. All data generation was conducted using the R software (R Development Core Team, 2008). Model estimation and goodness-of-fit computations were conducted with the flexMIRT® item response modeling software, version 2 (Cai, 2013b).

Insert Tables 1 and 2 about here

Insert Figure 2 about here

4.1 Calibration of the Test Statistic

Results for M_2 under the null conditions are shown in Table 3. Across the models evaluated—and for each sample size considered—the mean, variance, and empirical rejection rates obtained for the statistic across replications are close to what would be expected. Two-tailed Kolmogorov-Smirnov tests were used to evaluate the extent to which the observed distribution of M_2 differed from the expected chi-square reference distribution. At the $\alpha = 0.05$ level, the Kolmogorov-Smirnov test was not significant under any of the null conditions. The quantile-quantile plots in Figure 3 also show good match between the observed and expected distributions of the test statistic.

Insert Table 3 about here

Insert Figure 3 about here

Figure 4 shows a histogram of the empirical rejection rates for the Chen and Thissen (1997) $\mathrm{X}^2_{\mathrm{LD}}$ statistic, across the 276 item pairs. These rejection rates were obtained by tallying the number of times $\mathrm{X}^2_{\mathrm{LD}}$ exceeded the critical value for $\alpha = 0.05$ and a chi-squared distribution with one degree of freedom. It seems that the rejection rate is generally below the nominal level, averaging between 0.025 and 0.028 for the null conditions. This result is consistent with results obtained with the statistic in item response theory models (Chen & Thissen, 1997; Liu & Maydeu-Olivares, 2013), where $\mathrm{X}^2_{\mathrm{LD}}$ has been found to be somewhat conservative for dichotomous data. Although this may reduce the sensitivity of the index for

making judgments of the statistical significance of local dependence for a single item pair, in practice we often use the index to evaluate the relative severity of local independence violations across the many item pairs, so conservativeness may in fact be a welcoming feature under multiplicity of testing.

Insert Figure 4 about here

4.2 Power to Detect Unmodeled Testlets

The previous section demonstrated that M_2 had very good Type I error control when a hierarchical model was fit to data generated with testlets. Here, we present results obtained when a standard higher-order diagnostic model—that is, a model that ignores the testlet structure of the data generating model—is fit to the same data. As shown in Table 4, this model was rejected for every replication and at all α levels, regardless of the magnitude of the testlet effect or the sample size. It seems, then, that M_2 is quite sensitive to this type of model misspecification.

Insert Table 4 about here

Figure 5 provides heat maps depicting the average X^2_{LD} value (across replications) for each item pair. Darker values indicate larger values of the statistic. The patterns of local dependence revealed through X^2_{LD} are consistent with the known structure of the generating model. The fitted model does not adequately explain the strong covariation among items within each testlet, resulting in rather severe local dependence among these items.

Insert Figure 5 about here

4.3 Power to Detect Misspecifications of the Higher-order Structure

Results presented in this section were obtained by fitting a standard higher-order DINA model (i.e., one in which the higher-order dimension is univariate normal) to data generated from models with non-standard higher-order structures. The data generating models included higher-order latent dimension distributions that were non-normal or bivariate normal. Table 5 presents the M_2 empirical rejection rates for these conditions of model misspecification.

Insert Table 5 about here

Empirical rejection rates for the non-normal generating conditions are quite similar to the nominal α levels, indicating that M_2 is not sensitive to this type of misspecification. This result is consistent with a previous study examining the power of M_2 to detect non-normality in item response theory models (Li & Cai, 2012).

The results for the data generated from models with two higher-order dimensions varied according to the correlation between these dimensions. Specifically, rejection rates

increased as the correlation decreased (that is, as the higher-order structure of the generating model became less similar to the undimensional structure of the fitted model). When the correlation was $\rho = 0.8$, the rejection rates were only slightly elevated above the nominal α levels. Power was higher when $\rho = 0.4$, though still fairly low except for the largest sample size examined.

4.4 Power to Detect Misspecifications of the Q-matrix or the Item Model Type

Table 6 presents results for conditions in which there were errors in the Q-matrix of the fitted model or misspecification of the item model type. As described earlier, the Q-matrix errors included the addition or omission of paths (i.e., changing the values of individual Q-matrix elements, from 0 to 1 or vice versa) and the addition or omission of latent attributes (i.e., adding or deleting a column from the Q-matrix). For errors of model type, a compensatory (C-RUM) model was specified for either one item or for all items (the data were generated according to a DINA model for all items).

Insert Table 6 about here

The M_2 statistic was quite sensitive to three of the four Q-matrix specifications examined. There was good power to detect the addition or omission of paths, as well as to detect the omission of a latent attribute. In contrast, specification of extraneous attribute appeared to have no effect on the rejection rates, which were similar to the nominal α levels. Figure 6 shows the patterns of local dependence for these conditions, as reflected in the average X^2_{LD} values. The X^2_{LD} statistic clearly identified items with incorrect Q-matrix misspecifications, so long as the number of such misspecifications is small (the pattern of local dependence resulting from the omission of an attribute—which affects a much larger number of variables—would seem to be less interpretable).

Insert Figure 6 about here

When an extraneous attribute is included in the model, neither M_2 nor X^2_{LD} provide evidence of misspecification. However, inspection of the marginal attribute probabilities reveals that the expected probability of possessing the extraneous attribute is very close to 1. In a DINA model, then, it seems that such a variable can be absorbed without any change in model fit. In the example here, the extraneous attribute (x_5) nearly always takes a value of 1. In those cases, the third-order interaction is the same as the corresponding second order interaction (that is, if $x_5 = 1$, then $x_j x_{j'} x_5 = x_j x_{j'}$). As a result, the parameters estimated end up representing the same quantity ($\hat{\lambda}_{j \times j' \times 5} \approx \hat{\lambda}_{j \times j'}$).

The M_2 statistic appears to have some sensitivity to the misspecification of item type. However, when the error is limited to a single item, rejection rates are only slightly above the

nominal rates. Power is substantially higher when this misspecification is applied to all items in the test. For the conditions examined, X^2_{LD} does not appear to be particularly informative, as shown in Figure 7.

Insert Figure 7 about here

5 Analysis of Empirical Data

The results of the simulation study presented in the previous section suggest that fit statistics based on univariate and bivariate marginal subtables can be useful in identifying and characterizing model misfit. Here, we apply the proposed approach to an empirical example, using M_2 and the Chen and Thissen (1997) X^2_{LD} statistic to evaluate the fit of alternative diagnostic models to data from the 2007 Trends in Mathematics and Science Study (TIMSS). This example builds on prior work by Lee, Park, & Taylan (2011), who analyzed data from booklets 4 and 5 from the TIMSS 2007 fourth grade mathematics test. As part of their study, several teachers and content experts reviewed and coded the TIMSS test items according to the specific testing objectives described in the TIMSS 2007 framework. For the 25 items considered in the study, 15 unique testing objectives were identified (out of the 32 total objectives in the test framework). Accordingly, the Q-matrix proposed by Lee et al. (2011) consists of 25 rows (one for each item) and 15 columns (one for each attribute).

There is a good deal of variation in the number of items measuring each attribute. Ten of the 15 objectives are measured by only two or three items, while the second and third attributes are measured by 16 and 11 items, respectively. Among the 25 items, the number of underlying attributes ranges from 1 (items 2, 9, 24) to 6 (item 14).

Lee et al. (2011) specified a conjunctive (DINA) model for the items in their analysis. For the current study, we initially fit a higher-order version of this model (de la Torre & Douglas, 2004) using the Q-matrix exactly as reported in the earlier study to a sample of 564 students from the United States. As shown in the first row of Table 7, the value of M_2 for this model was 391.0, with 259 degrees of freedom. The RMSEA based on M_2 has a value of 0.030, with 90% confidence interval of (0.000, 0.036).

Insert Table 7 about here

Values of the Chen and Thissen (1997) X^2_{LD} statistic are presented for this model in the left panel of Figure 8. There are a handful of item pairs with fairly large X^2_{LD} values, indicating substantial local dependence. For illustrative purposes, we focused our attention here on the two item pairs with the largest X^2_{LD} values: items 1—5 ($X^2_{LD} = 13.8$) and items 18—19 ($X^2_{LD} = 38.0$). Table 8 presents the observed marginal response proportions and model-implied probabilities from which the test statistic was computed.

Insert Figure 8 about here

Insert Table 8 about here

We examined the two item pairs and their specification within the initial fitted model for potential causes of the observed local dependence. The items considered in this study have all been released (Foy & Olson, 2009) and were thus available for review. Items 18 and 19 are administered as within a cluster or testlet (M031242A/B/C), and it was evident upon inspecting the items that a correct response to item 18 (M031242A) would seem to greatly simplify the task of answering item 19; the answer to the question 19 (M031242C) can quite simply be read from a table that the examinee is asked to complete for item 18. This may explain why the diagnostic model does not fully explain the covariation in responses between these two items. Although it would be possible to arrive at the correct answer to item 19 by applying the skills identified in the Q-matrix as being relevant, those skills are less necessary once an examinee answers item 18. In order to model this lack of independence between these items (conditional on the attributes), a testlet effect could be specified. It should be noted that item 20 (M031242C) is also part of the same item cluster. However, this item does not rely quite as directly on information from items 18 or 19, so we chose to ignore local dependence between item 20 and items 18 and 19. Similarly, items 9 and 10 (M041258A and M041258B, respectively) are administered as a testlet but show very little evidence of local dependence, so no random effect is specified for this pair.

Our review of items 1 and 5 identified two possible changes in the Q-matrix. First, although item 1 (M041052) had been described in the study by Lee et al. (2011) as requiring both attributes 1 and 2, it seemed to us that possession of attribute 1 would be sufficient to obtain the correct answer. Thus, we posited an alternative Q-matrix specification for item 1, with this item depending only on possession of attribute 1. After examination of item 5, we concluded that the specification of dependence on attributes 2 and 3 was reasonable. However, the relevance of attribute 8 was unclear. Thus, a second Q-matrix change was considered, with item 5 depending on attributes 2 and 3 but not attribute 8.

In summary, our alternative model for the TIMSS data included three changes. Two changes were made to the Q-matrix: the values of elements $q_{1,2}$ and $q_{5,8}$ were changed from 1 to 0. In addition, a testlet effect was added to account for the strong dependence between items 18 and 19. The total number of estimated parameters is 67 for the alternative model—one more than required for the initial model. The Q-matrix changes for items 1 and 5 do not affect the number of free parameters. For item 1, the interaction $\gamma_{1,1\times2}$ is fixed to 0, but the main effect $\gamma_{1,1}$ is now estimated. Similarly, for item 5, the third-order interaction $\gamma_{5,2\times3\times8}$ is fixed, and $\gamma_{5,2\times3}$is estimated. The one additional parameter estimated in the alternative model

is the slope parameter for the testlet effect (for identification, the slopes of items 18 and 19 were constrained to be equal—i.e., $\beta_{18,1} = \beta_{19,1} = \beta$).

Overall goodness-of-fit indices for the alternative model are presented in the second row of Table 7. The value of M_2 for the alternative model was 330.3, with 258 degrees of freedom. The RMSEA based on M_2 has a value of 0.022, with 90% confidence interval of (0.000, 0.029). Results for the Chen & Thissen (1997) X^2_{LD} are shown in the second row of Table 8 and in the right panel of Figure 8. The fit of the alternative model is slightly improved over the initial model, both on the basis of the limited-information fit statistics and the information-based indices (AIC and BIC). It should be noted, however, that this improvement is rather modest. In addition, while specification of the testlet effect seems to have fully accounted for the local dependence of items 18 and 19 (with X^2_{LD} decreased from 38.0 to 0.9), there remains some dependence between items 1 and 5 (X^2_{LD} only decreased from 13.8 to 11.0), despite the changes in model specification.

Of course, for this empirical study, we cannot know the true generating model. That said, our analyses are primarily intended to illustrate the ways in which researchers might use the goodness-of-fit statistics to evaluate the fit of models, to characterize possible misspecifications, and to identify candidate alternative models (and then test these alternatives, though ideally this would be done with an independent sample of respondents).

6 Discussion

In this paper, we demonstrate the application of limited-information fit statistics to diagnostic classification models. Through simulation studies we found that M_2 is well calibrated, closely matching its reference distribution. This result was observed across a wide range of conditions, including standard higher-order diagnostic classification models, hierarchical models (e.g., with testlet effects), and models with correlated higher-order dimensions.

We also examined the sensitivity of M_2 to a number of different kinds of model misspecification, including (1) failure to account for testlet-type effects, (2) incorrect specification of higher-order distribution or structure (fitting higher-order models with univariate normal distributions of the higher-order dimension to data generated from non-normal or bivariate normal distributions, (3) errors in the Q-matrix, and (4) misspecifications of item type (C-RUM instead of DINA). The results here were mixed. M_2 was found to be highly sensitive to the presence of testlet effects and certain Q-matrix misspecifications (addition or omission of paths, omission of an attribute). In contrast, M_2 was largely insensitive to misspecifications of the higher-order structure. Misfit was not detected when the higher-order dimension was sampled from any of the non-normal distributions examined.

When the generating model included correlated higher-order dimensions, power was relatively low for strongly correlated dimensions but increased as the correlation decreased.

There are of course, many limitations to the work presented here. First, we have focused on the application of M_2 to dichotomous item response data. However, it would be useful to examine the performance of the test statistic with polytomous models. Prior work has suggested that even bivariate marginal subtables may be poorly filled as the number of categories increases, potentially reducing the utility of M_2 (Cai & Hansen, 2013). Various methods for collapsing the subtables have been proposed, and seem to perform well in applications of item response theory (Joe & Maydeu-Olivares, 2010; Maydeu-Olivares, Cai, & Hernandez, 2011; Cai & Hansen, 2013; Cai & Monroe, 2014). However, it would be useful to evaluate these approaches in the context of diagnostic classification modeling.

A second limitation is that in our simulation study we held constant certain aspects of the data-generating model. For example, the number of attributes (4), the number of test items (24), the structure of the Q-matrix, and the item type (DINA) were the same across all conditions. Although this was done in order to put some bounds on the scope of this study, one might wonder whether the findings might differ if any of these features were allowed to vary.

It is noteworthy that the test statistics utilized here—M_2 and the Chen and Thissen (1997) X_{LD}^2—have recently been implemented for diagnostic classifications models in commercially item response modeling software (Cai, 2013b). We expect that increased availability of these tools for evaluating fit will further clarify their potential utility, as well as their limitations.

7 References

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2^p contingency tables. *British Journal of Mathematical and Statistical Psychology, 55,* 1–15.

Bartholomew, D. J., & Tzamourani, P. (1999). The goodness-of-fit of latent trait models in attitude measurement. *Sociological Methods and Research, 27,* 525–546.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete multivariate analysis: theory and practice.* Springer.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Cai, L. (2013a, October). *Flexible multidimensional item response theory analysis and model fit evaluation.* Cattell award address presented at the 2013 meeting of the Society of Multivariate Experimental Psychology. St. Pete Beach, FL.

Cai, L. (2013b). *flexMIRT® 2.0: Flexible multilevel multidimensional item analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L. & Monroe, S. (2014). *A new statistic for evaluating item response theory models for ordinal data.* (CRESST Report 839). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66,* 245-276.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited- information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology, 59,* 173–194.

Cai, L., Yang, J., & Hansen, M. (2011). Generalized full information bifactor analysis. *Psychological Methods, 16,* 221-248.

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

Choi, H.-J., Rupp, A. A., & Pan, M. (2013). Standardized diagnostic assessment design and analysis: key ideas from modern measurement theory. In M. M. C. Mok (Ed.), *Self-directed learning oriented assessments in the Asia-Pacific, Education in the Asia-Pacific region: Issues, Concerns and Prospects 18* (pp. 61–85). Dordrecht: Springer.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: development and applications. *Journal of Educational Measurement, 45,* 343–362.

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

Foy, P., & Olson, J. F. (Eds.). (2009). *TIMSS 2007 user guide for the international database.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bifactor analysis. *Psychometrika, 57,* 423–436.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign.

Henson, R., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74,* 191–210.

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75,* 393–419.

Junker, B., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Jurich, D. P., Bradshaw, L. P., DeMars, C. E. (2014, April). *Limited-information methods to assess overall fit of diagnostic classification models.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Philadelphia, PA.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement, 49,* 59–81.

Lai, H., Cui, Y., & Gierl, M. J. (2012, April). *Item consistency index: an item-fit index for cognitive diagnostic assessment.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Vancouver, BC, Canada.

Lee, Y., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic models of at- tribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing, 11,* 144–177.

Li, Z., Cai, L. (2012, July). *Summed score based fit indices for testing latent variable distribution assumption in IRT*. Paper presented at the 2012 International Meeting of the Psychometric Society, Lincoln, NE.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement, 73,* 254-274.

Maris, G., & Bechger, T. (2009). Equivalent diagnostic classification models. *Measurement: Interdisciplinary Research and Perspectives, 7,* 41-46.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11,* 71-101.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713–732.

Maydeu-Olivares, A., Cai, L., & Hernandez, A. (2011). Comparing the fit of IRT and factor analysis models. *Structural Equation Modeling, 18,* 333–356.

R Development Core Team (2008). *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Rijmen, F. (2009). *Efficient full-information maximum likelihood estimation for multidimensional IRT models.* (Technical Report No. RR-09-03). Princeton, NJ: Educational Testing Service.

Rupp, A. A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68,* 78-96.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications.* New York: Guilford.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: a case study. *Educational and Psychological Measurement, 2,* 239-257.

Templin, J. (2007, October). *Assessing cognitive diagnosis model fit using limited information methods.* Paper presented at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics in Greensboro, North Carolina.

Templin, J. L., & Henson, R. A. (2010). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 37,* 287-305.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11,* 287–305.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS No. RR-05-16). Princeton, NJ: ETS.

Wilhelm, O., & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. *Measurement: Interdisciplinary Research and Perspectives, 7,* 53-57.

Woods, C. M., & Lin, N. (2009). Item response theory with estimation of the latent density using Davidian curves. *Applied Psychological Measurement, 33,* 102-117.

8 Tables

Table 1

Data Generation for Simulation Study: Q-Matrix and Item Parameters

i	Q-matrix				α_i	Attribute Two-Way Interaction Terms						Testlet Slope Parameters					
	$q_{i,1}$	$q_{i,2}$	$q_{i,3}$	$q_{i,4}$		$\gamma_{i,1x2}$	$\gamma_{i,1x3}$	$\gamma_{i,1x4}$	$\gamma_{i,2x3}$	$\gamma_{i,2x4}$	$\gamma_{i,3x4}$	$\beta_{i,1}$	$\beta_{i,2}$	$\beta_{i,3}$	$\beta_{i,4}$	$\beta_{i,5}$	$\beta_{i,6}$
1	0	1	1	0	-1.29				3.76			(0,1,2)					
2	1	1	0	0	-1.04	3.00						(0,1,2)					
3	0	1	1	0	-1.36				2.90			(0,1,2)					
4	1	0	1	0	-1.00		2.06					(0,1,2)					
5	1	0	1	0	-1.36		3.14						(0,1,2)				
6	0	1	0	1	-.95					2.51			(0,1,2)				
7	0	0	1	1	-.87						2.63		(0,1,2)				
8	0	0	1	1	-1.19						3.61		(0,1,2)				
9	0	1	1	0	-.59				2.52					(0,1,2)			
10	0	1	0	1	-.98					2.90				(0,1,2)			
11	0	1	0	1	-1.29					3.05				(0,1,2)			
12	0	1	0	1	-.74					1.94				(0,1,2)			
13	1	0	1	0	-1.11		2.92								(0,1,2)		
14	1	0	0	1	-1.10			2.95							(0,1,2)		
15	1	1	0	0	-.80	2.21		.00							(0,1,2)		
16	1	0	0	1	-.84			2.32							(0,1,2)		
17	1	0	0	1	-.92			2.97								(0,1,2)	
18	1	0	0	1	-.81			2.87								(0,1,2)	
19	1	0	1	0	-.85		2.42									(0,1,2)	
20	1	1	0	0	-1.08	2.45										(0,1,2)	
21	0	0	1	1	-1.81						3.60						(0,1,2)
22	1	1	0	0	-.91	2.23											(0,1,2)
23	0	0	1	1	-1.12						2.67						(0,1,2)
24	0	1	1	0	-1.15				3.17								(0,1,2)

Note: In generating model (DINA), the attribute main effects (which are not shown) are fixed to zero.

Table 2

Summary of data generating and fitted models used in simulation study

Data Generating Model			Fitted Model		
Higher-Order Structure	Attribute Model	Testlets	Higher-Order Structure	Attribute Model	Testlet
null (correctly specified) models (results in Table 3)					
univariate normal	DINA	no	univariate normal	DINA	no
univariate normal	DINA	yes ($\beta = 1$)	univariate normal	DINA	yes
univariate normal	DINA	yes ($\beta = 2$)	univariate normal	DINA	yes
bivariate normal ($\rho = 0.4$)	DINA	no	bivariate normal	DINA	no
bivariate normal ($\rho = 0.6$)	DINA	no	bivariate normal	DINA	no
bivariate normal ($\rho = 0.8$)	DINA	no	bivariate normal	DINA	no
failure to model testlet effects (results in Table 4)					
univariate normal	DINA	yes ($\beta = 1$)	univariate normal	DINA	no
univariate normal	DINA	yes ($\beta = 2$)	univariate normal	DINA	no
misspecifications of higher-order latent variable distributions (results in Table 5)					
univariate bimodal	DINA	no	univariate normal	DINA	no
univariate extreme bimodal	DINA	no	univariate normal	DINA	no
univariate right-skewed	DINA	no	univariate normal	DINA	no
univariate extreme right skewed	DINA	no	univariate normal	DINA	no
bivariate normal ($\rho = 0.4$)	DINA	no	univariate normal	DINA	no
bivariate normal ($\rho = 0.6$)	DINA	no	univariate normal	DINA	no
bivariate normal ($\rho = 0.8$)	DINA	no	univariate normal	DINA	no
Q-matrix or item type misspecifications (results in Table 6)					
univariate normal	DINA	no	univariate normal	C-RUM (item 8)	no
univariate normal	DINA	no	univariate normal	C-RUM (all items)	no
univariate normal	DINA	no	univariate normal	omit path	no
univariate normal	DINA	no	univariate normal	add path	no
univariate normal	DINA	no	univariate normal	omit attribute	no
univariate normal	DINA	no	univariate normal	add attribute	no

Table 3

Simulation Study Results: M_2 Calibration under Null Conditions for Various Data Generating Models

N	reps	df	M	V	KS	Empirical Rejection Rate				
						.200	.150	.100	.050	.010
higher-order DINA										
500	500	247	248.4	518.4	.356	.206	.162	.136	.064	.014
1000	500	247	248.0	508.7	.757	.212	.154	.104	.060	.018
2000	500	247	247.0	524.3	.692	.212	.146	.108	.054	.014
higher-order DINA with testlet effects, testlet slope $\beta = 1$										
500	500	241	240.7	450.9	.931	.200	.140	.078	.040	.002
1000	500	241	240.0	436.5	.787	.186	.128	.076	.036	.006
2000	500	241	240.4	499.5	.198	.198	.156	.116	.050	.004
higher-order DINA with testlet effects, testlet slope $\beta = 2$										
500	499	241	242.2	459.7	.311	.212	.166	.120	.062	.014
1000	500	241	242.6	450.0	.049	.204	.142	.086	.056	.008
2000	500	241	240.3	435.2	.352	.186	.134	.090	.042	.004
DINA with bivariate normal higher-order latent distribution, $\rho = 0.4$										
500	496	245	245.0	480.8	.913	.179	.139	.101	.050	.008
1000	499	245	246.3	503.0	.146	.220	.160	.106	.062	.010
2000	500	245	243.3	488.5	.079	.186	.140	.078	.040	.016
DINA with bivariate normal higher-order latent distribution, $\rho = 0.6$										
500	483	245	245.5	486.6	.551	.199	.147	.099	.050	.012
1000	498	245	246.6	494.8	.057	.219	.173	.100	.056	.008
2000	500	245	244.0	480.6	.557	.172	.122	.088	.046	.012
DINA with bivariate normal higher-order latent distribution, $\rho = 0.8$										
500	394	245	244.2	473.6	.953	.198	.140	.091	.038	.003
1000	464	245	244.2	480.4	.805	.196	.157	.093	.039	.002
2000	492	245	244.1	481.9	.390	.195	.146	.093	.051	.008

Note: N is the sample size; *reps* is the number of converged replications (out of 500); *df* is the degrees of freedom for M_2, given the fitted model; M and V observed the mean and variance, respectively, of M_2 across the converged replications within the condition; KS indicates the *p*-values for two-tailed Kolmogorov-Smirnov test.

Table 4

Simulation Study Results: M_2 Power to Detect Testlet Effects

N	reps	df	Empirical Rejection Rate					RMSEA	
			.200	.150	.100	.050	.010	M	(90% CI)
higher-order DINA with testlet effects, testlet slope $\beta = 1$									
500	500	247	1.000	1.000	1.000	1.000	1.000	.044	(.036,.051)
1000	500	247	1.000	1.000	1.000	1.000	1.000	.044	(.040,.048)
2000	500	247	1.000	1.000	1.000	1.000	1.000	.044	(.041,.047)
higher-order DINA with testlet effects, testlet slope $\beta = 2$									
500	500	247	1.000	1.000	1.000	1.000	1.000	.114	(.107,.120)
1000	500	247	1.000	1.000	1.000	1.000	1.000	.115	(.110,.120)
2000	500	247	1.000	1.000	1.000	1.000	1.000	.115	(.111,.118)

Note: *N* is the sample size; *reps* is the number of converged replications (out of 500); *df* is the degrees of freedom for M_2, given the fitted model.

Table 5

Simulation Study Results: M_2 Power to Detect Misspecifications of Higher-Order Latent Variable Distributions

N	reps	df	Empirical Rejection Rate					RMSEA	
			.200	.150	.100	.050	.010	M	(90% CI)
DINA with bimodal higher-order latent variable distribution									
500	500	247	.184	.134	.090	.044	.006	.006	(.000,.017)
1000	500	247	.180	.130	.090	.038	.006	.004	(.000,.012)
2000	500	247	.216	.168	.100	.056	.010	.003	(.000,.009)
DINA with extreme bimodal higher-order latent variable distribution									
500	500	247	.220	.156	.096	.058	.012	.006	(.000,.018)
1000	500	247	.206	.160	.120	.048	.014	.004	(.000,.012)
2000	500	247	.220	.176	.116	.052	.008	.003	(.000,.009)
DINA with right-skewed higher-order latent variable distribution									
500	500	247	.188	.128	.086	.028	.006	.005	(.000,.017)
1000	500	247	.204	.164	.108	.050	.018	.004	(.000,.012)
2000	500	247	.216	.160	.110	.038	.008	.003	(.000,.009)
DINA with extreme right-skewed higher-order latent variable distribution									
500	500	247	.196	.142	.102	.060	.010	.006	(.000,.018)
1000	500	247	.242	.176	.124	.062	.008	.004	(.000,.013)
2000	500	247	.190	.130	.086	.040	.004	.002	(.000,.008)
DINA with bivariate normal higher-order latent distribution, $\rho = 0.4$									
500	500	247	.290	.224	.158	.088	.022	.007	(.000,.019)
1000	500	247	.452	.364	.290	.164	.056	.007	(.000,.015)
2000	500	247	.626	.536	.450	.312	.134	.007	(.000,.012)
DINA with bivariate normal higher-order latent distribution, $\rho = 0.6$									
500	500	247	.246	.190	.124	.064	.022	.006	(.000,.018)
1000	500	247	.330	.252	.174	.092	.026	.006	(.000,.014)
2000	500	247	.416	.332	.246	.150	.036	.005	(.000,.010)
DINA with bivariate normal higher-order latent distribution, $\rho = 0.8$									
500	500	247	.236	.170	.100	.050	.014	.006	(.000,.017)
1000	500	247	.240	.192	.114	.056	.004	.004	(.000,.012)
2000	500	247	.238	.184	.130	.066	.014	.003	(.000,.009)

Note: *N* is the sample size; *reps* is the number of converged replications (out of 500); *df* is the degrees of freedom for M_2, given the fitted model.

Table 6

Simulation Study Results: M_2 Power to Detect to Q-Matrix or Item Type Misspecifications

N	reps	df	Empirical Rejection Rate					RMSEA	
			.200	.150	.100	.050	.010	M	(90% CI)
omit paths from attribute to items ($x_1 \rightarrow y_5$, $x_1 \rightarrow y_{16}$)									
500	500	247	.978	.968	.940	.888	.730	.024	(.015,.032)
1000	500	247	1.000	1.000	1.000	1.000	.998	.024	(.019,.029)
2000	500	247	1.000	1.000	1.000	1.000	1.000	.025	(.021,.028)
add (extraneous) paths from attribute to items ($x_1 \rightarrow y_3$, $x_1 \rightarrow y_{23}$)									
500	500	247	.920	.892	.850	.772	.566	.021	(.010,.029)
1000	500	247	1.000	1.000	1.000	1.000	.984	.022	(.017,.027)
2000	500	247	1.000	1.000	1.000	1.000	1.000	.022	(.019,.025)
omit attribute (x_4)									
500	500	248	.472	.392	.306	.206	.064	.010	(.000,.021)
1000	500	248	.738	.680	.590	.452	.202	.011	(.000,.019)
2000	500	248	.974	.956	.940	.888	.740	.012	(.007,.016)
add (extraneous) attribute (x_5)									
500	500	246	.204	.166	.138	.058	.012	.006	(.000,.018)
1000	500	246	.206	.154	.108	.056	.018	.004	(.000,.012)
2000	499	246	.206	.148	.100	.054	.012	.003	(.000,.009)
apply an incorrect item type (C-RUM for item 8)									
500	500	246	.230	.180	.138	.066	.018	.006	(.000,.018)
1000	500	246	.246	.186	.128	.074	.018	.004	(.000,.013)
2000	500	246	.280	.208	.144	.076	.024	.003	(.000,.009)
apply an incorrect item type (C-RUM for all items)									
500	500	223	.966	.952	.932	.866	.710	.025	(.014,.034)
1000	500	223	1.000	1.000	1.000	1.000	.990	.025	(.020,.031)
2000	500	223	1.000	1.000	1.000	1.000	1.000	.026	(.022,.029)

Note: *N* is the sample size; *reps* is the number of converged replications (out of 500); *df* is the degrees of freedom for M_2, given the fitted model.

Table 7

Empirical Illustration: Overall Goodness-of-Fit Evaluation for the Two Fitted Models.

	df	M_2	*p*	RMSEA	(90% C.I.)	AIC	BIC
model 1	259	391.0	<.001	.030	(.000, .036)	15872.4	16158.5
model 2	258	330.3	.002	.022	(.000, .029)	15821.3	16111.8

Table 8

Empirical Illustration: Bivariate Marginal Response Pattern Observed Proportions and Model-Implied Probabilities for 2 Item Pairs (1—5 and 18—19) and Corresponding X^2_{LD} Values for the Two Fitted Models

	Observed				Model-Implied				X^2_{LD}	*p*
	p_{00}	p_{01}	p_{10}	p_{11}	$\hat{\pi}_{00}$	$\hat{\pi}_{01}$	$\hat{\pi}_{10}$	$\hat{\pi}_{11}$		
items 1,5										
model 1	.128	.073	.220	.580	.097	.103	.252	.549	13.8	< .001
model 2					.103	.101	.250	.546	11.0	< .001
items 18,29										
model 1	.314	.046	.232	.408	.257	.104	.291	.348	38.0	< .001
model 2					.310	.053	.241	.396	0.9	.331

9 Figure Captions

Figure 1. Densities from which scores on the higher-order dimension were sampled for non-normal data generating conditions.

Figure 2. Path diagrams for data generating models used in simulation study.

Figure 3. Simulation study results: Quantile–quantile plots of observed M_2 values and their reference chi-square distributions (degrees of freedom shown in the subscripts of the x-axis labels). Closed grey circles indicate results for conditions with sample size N = 500, open black circles indicate results for N = 1000, and plus signs indicate N = 2000. Reported p-values are for a two-tailed Kolmogorov–Smirnov test of the equality of the observed M_2 distributions with its corresponding reference distribution.

Figure 4. Simulation study results: Empirical rejection rates for Chen and Thissen (1997) X_{LD}^2 across all item pairs. Results are shown for null conditions (fitted model correctly specified) with sample size N = 2000 and α = 0.05. Rejection rates are based on all converged replications (minimum of 492; maximum of 500).

Figure 5. Simulation study results: Chen and Thissen (1997) X_{LD}^2 heat map for diagnostic models with testlet effects. Shading is based on the average RMSEA for the item pair across all converged replications. The generating model in both panels is a higher-order DINA model with testlet effects. The testlet slope parameters are $\beta = 1$ and $\beta = 2$ for the left and right panels, respectively. The fitted models do not include the testlet effects.

Figure 6. Simulation study results: Chen and Thissen (1997) X_{LD}^2 heat map for diagnostic models with incorrect specification of item type. Shading is based on the average RMSEA for the item pair across all converged replications. The generating model is a higher-order DINA model. In left panel, item 8 (indicated by tick mark on the x- and y-axes) is incorrectly specified as C-RUM. In right panel, all 24 items are incorrectly specified as C-RUM.

Figure 7. Simulation study results: Chen and Thissen (1997) X_{LD}^2heat map for diagnostic models with Q-matrix misspecification. Shading is based on the average RMSEA for the item pair across all converged replications. The generating model in all four panels is a higher-order DINA model. In top-left panel, paths from x_1 to y_5 and from x_1 to y_{16} are omitted (i.e., the values of Q-matrix elements $q_{5,1}$ and $q_{16,1}$ were changed from 1 to 0; tickmarks identify items 5 and 16). In the top-right panel, extraneous paths were created from $x_1$to y_3 and from x_1 to y_{23} (i.e., the values of Q-matrix elements $q_{3,1}$ and $q_{23,1}$ were changed from 0 to 1; tickmarks identify items 3 and 23). In the bottom-left panel, attribute x_4 is omitted (i.e., all values in column 4 of the Q-matrix are 0; 12 elements that have values of 1 in the Q-matrix of the generating model are identified by tickmarks). In the bottom-right panel, an extraneous attribute, x_5 is specified (i.e., a 5th column is added to the Q-matrix); 4 items (2, 6, 8, and 13); indicated by tickmarks on the x- and y-axes) have Q-matrix values of 1 for this attribute.

Figure 8. Results from empirical illustration: Chen and Thissen (1997) X_{LD}^2 heat map for analysis of 25 items from TIMSS 4th grade math booklet 4. Left panel shows results obtained by fitting a higher-order DINA model with Q-matrix as described by Lee, Park, & Taylan (2011). Right panel shows results obtained by fitting a model with a slightly altered Q-matrix and the addition of a testlet effect (for items 18 and 19).

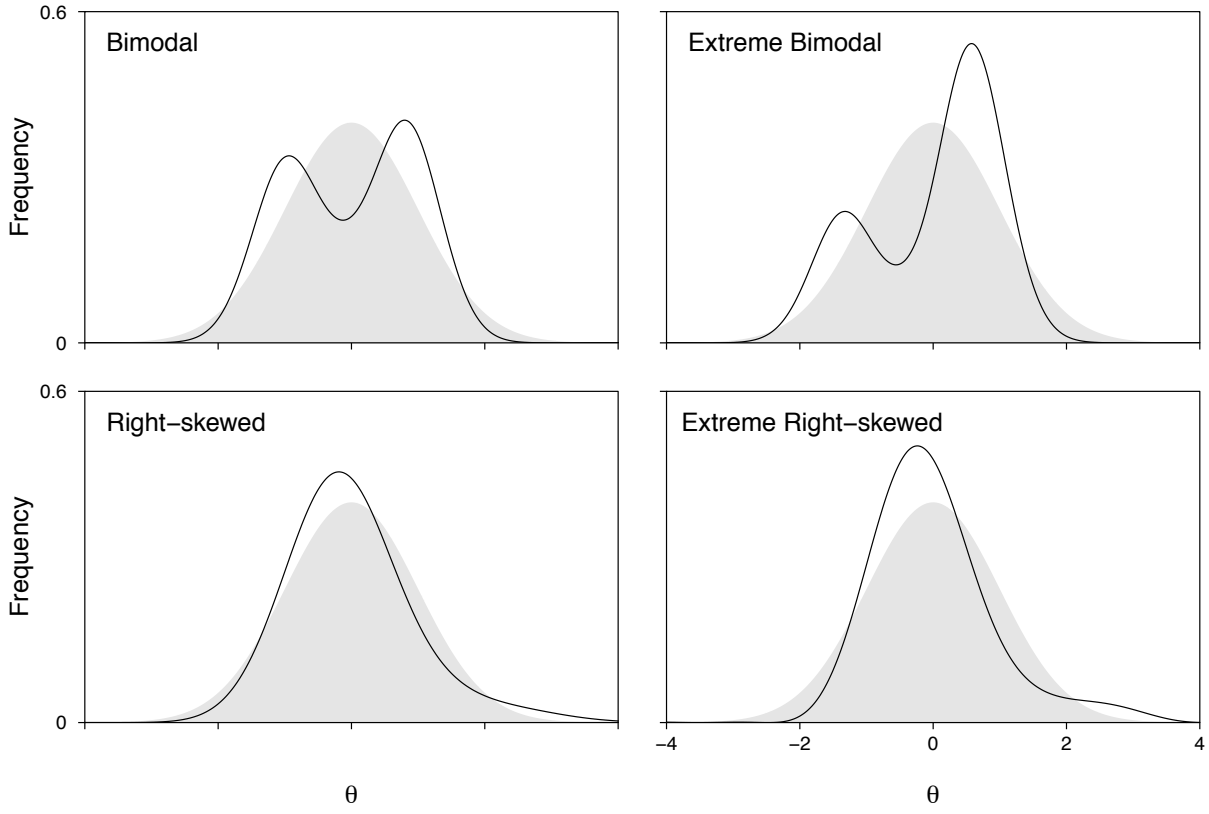
Bimodal
Extreme Bimodal
Right−skewed
Extreme Right−skewed
Frequency
Frequency
0.6
0
0.6
0
−4
−2
0
2
4
θ
θ

Figure 1

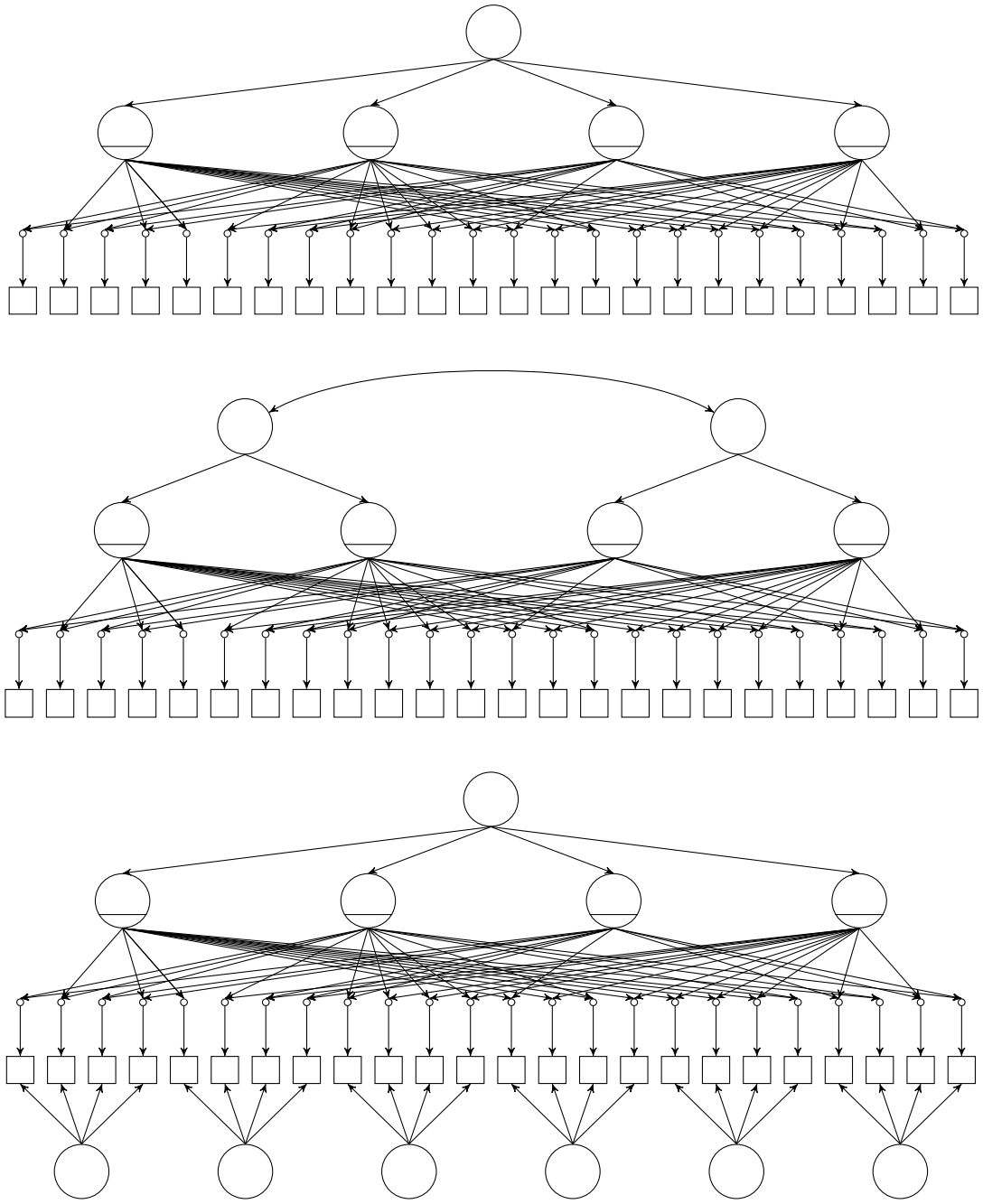

Figure 2

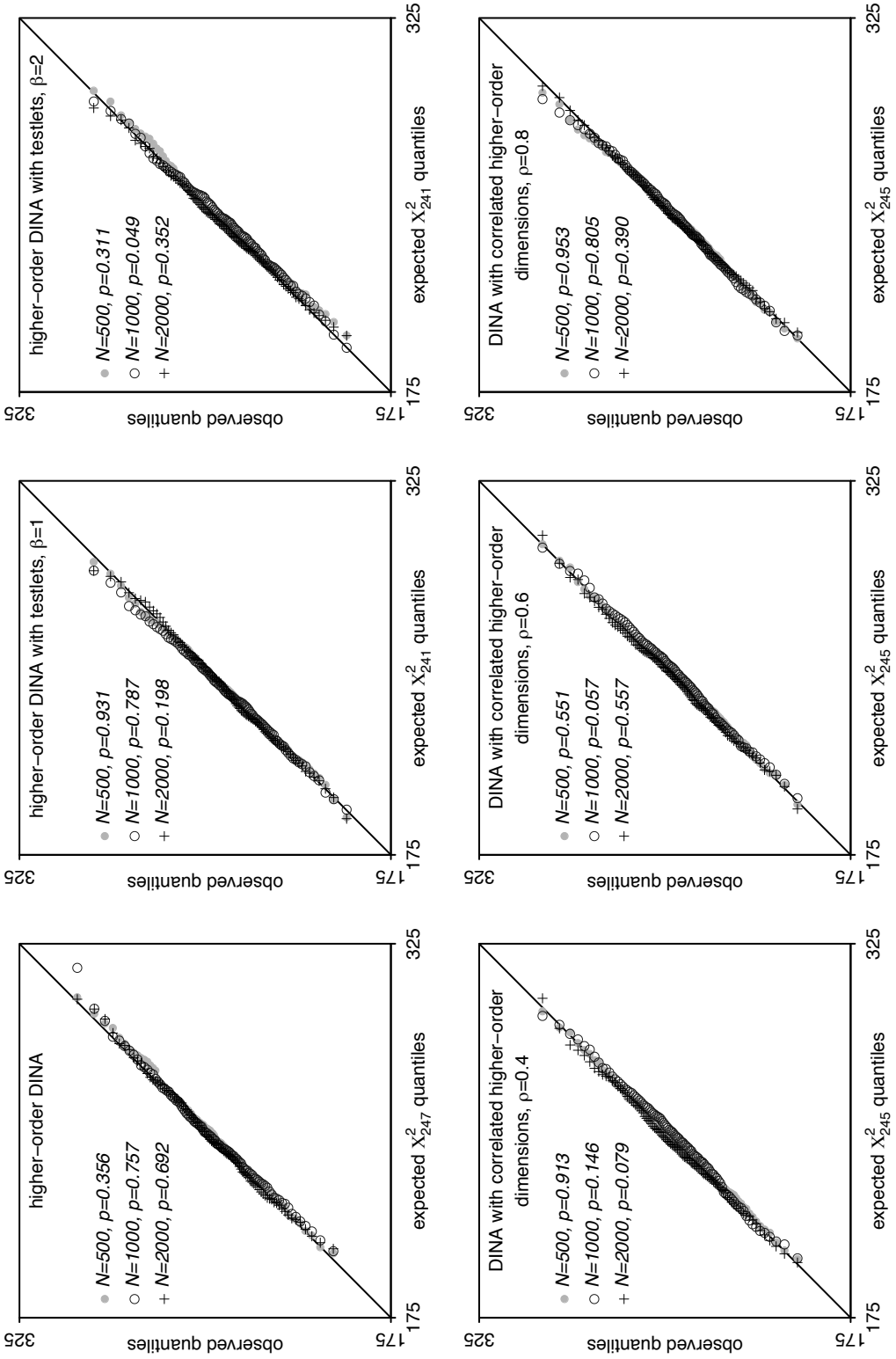
higher−order DINA
N=500, p=0.356
N=1000, p=0.757
N=2000, p=0.692
higher−order DINA with testlets, β=1
N=500, p=0.931
N=1000, p=0.787
N=2000, p=0.198
higher−order DINA with testlets, β=2
N=500, p=0.311
N=1000, p=0.049
N=2000, p=0.352
DINA with correlated higher−order dimensions, ρ=0.4
N=500, p=0.913
N=1000, p=0.146
N=2000, p=0.079
DINA with correlated higher−order dimensions, ρ=0.6
N=500, p=0.551
N=1000, p=0.057
N=2000, p=0.557
DINA with correlated higher−order dimensions, ρ=0.8
N=500, p=0.953
N=1000, p=0.805
N=2000, p=0.390
observed quantiles
expected X^2_{247} quantiles
expected X^2_{241} quantiles
expected X^2_{245} quantiles
175
325

Figure 3

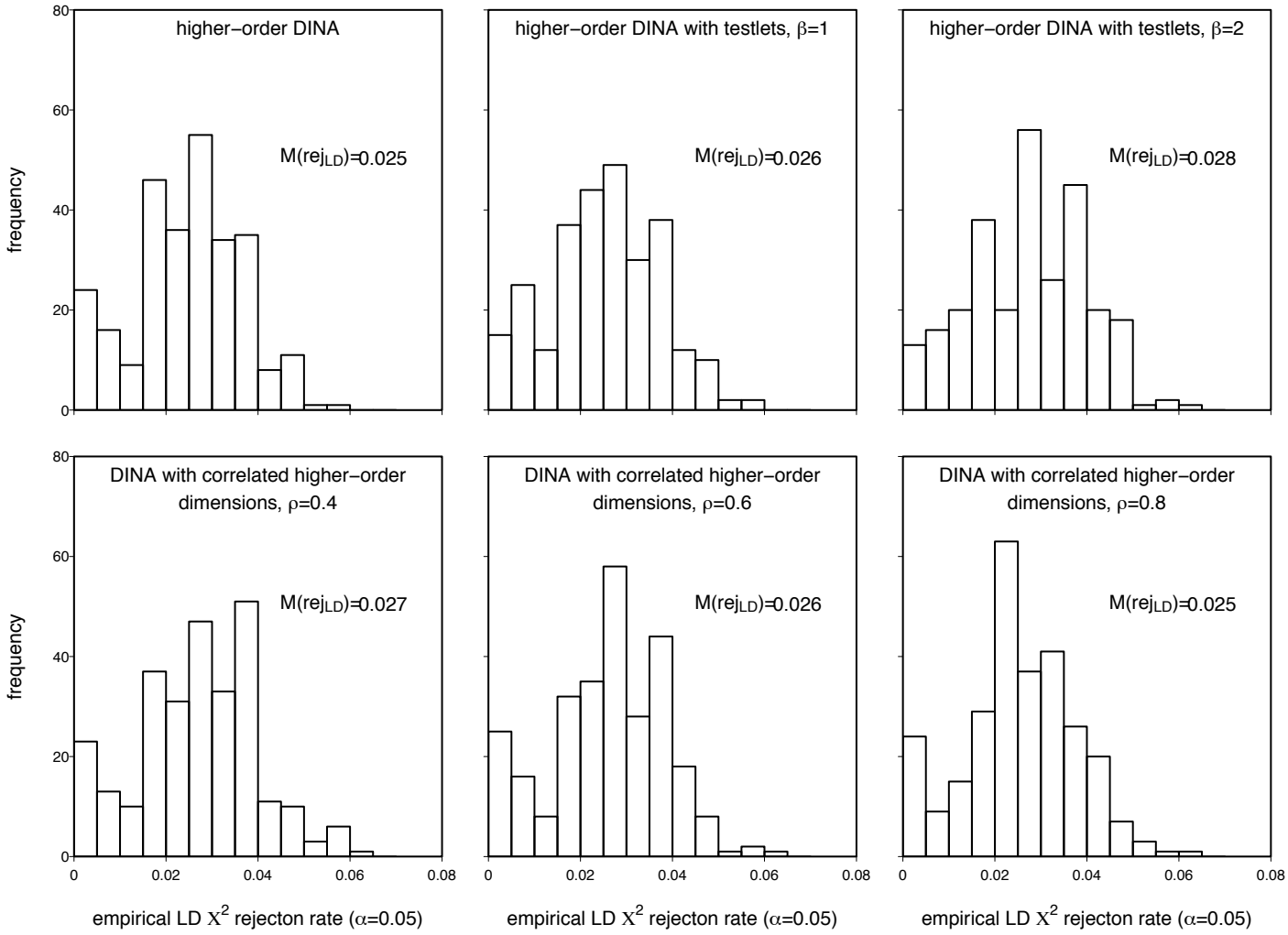
higher−order DINA
M(rej$_{LD}$)=0.025
higher−order DINA with testlets, β=1
M(rej$_{LD}$)=0.026
higher−order DINA with testlets, β=2
M(rej$_{LD}$)=0.028
DINA with correlated higher−order dimensions, ρ=0.4
M(rej$_{LD}$)=0.027
DINA with correlated higher−order dimensions, ρ=0.6
M(rej$_{LD}$)=0.026
DINA with correlated higher−order dimensions, ρ=0.8
M(rej$_{LD}$)=0.025
frequency
0
20
40
60
80
0
0.02
0.04
0.06
0.08
empirical LD X^2 rejecton rate (α=0.05)

Figure 4

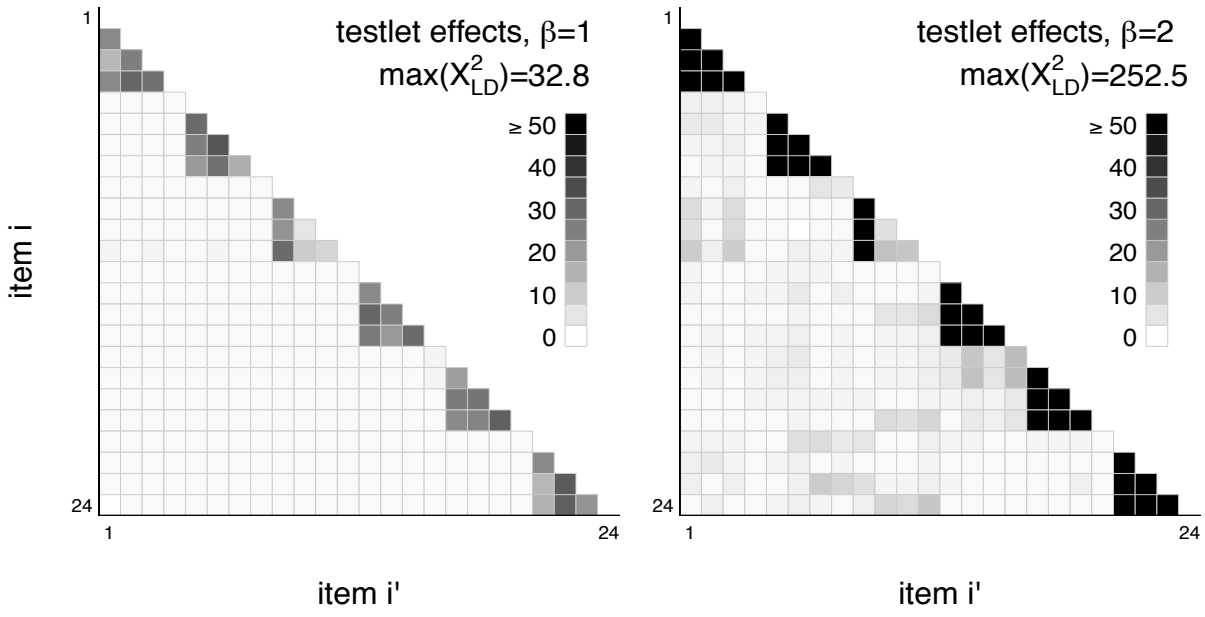
testlet effects, β=1
max(X^2_{LD})=32.8
testlet effects, β=2
max(X^2_{LD})=252.5
≥ 50
40
30
20
10
0
item i
item i'
1
24

Figure 5

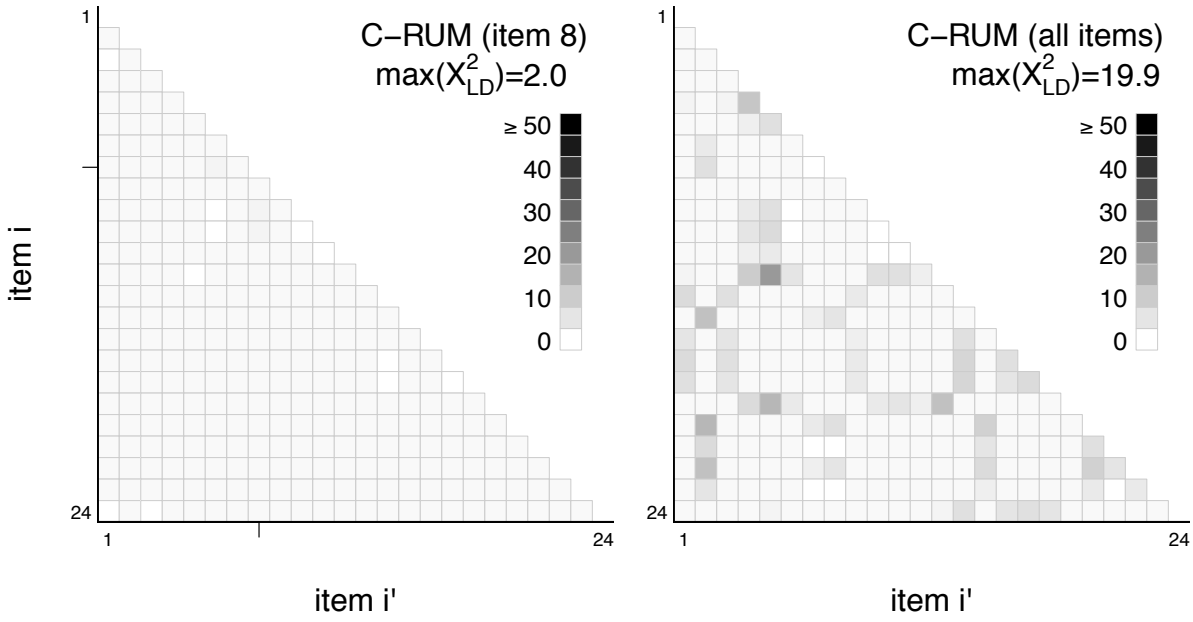
C−RUM (item 8)
max(X^2_{LD})=2.0
C−RUM (all items)
max(X^2_{LD})=19.9
≥ 50
40
30
20
10
0
item i
item i'
1
24

Figure 6

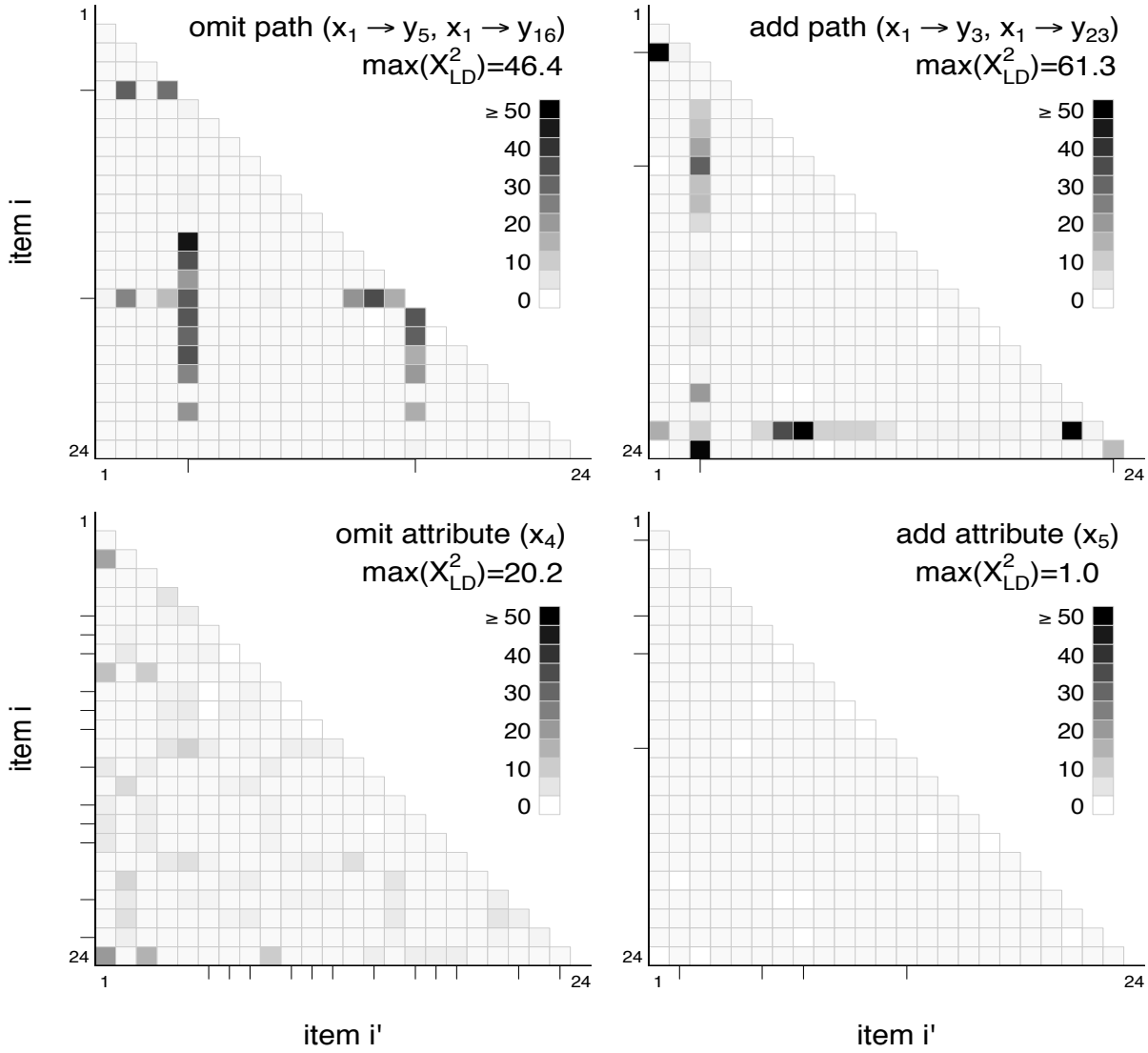
omit path ($x_1 \rightarrow y_5$, $x_1 \rightarrow y_{16}$)
$\max(X^2_{LD})$=46.4
add path ($x_1 \rightarrow y_3$, $x_1 \rightarrow y_{23}$)
$\max(X^2_{LD})$=61.3
omit attribute (x_4)
$\max(X^2_{LD})$=20.2
add attribute (x_5)
$\max(X^2_{LD})$=1.0
≥ 50
40
30
20
10
0
item i
item i'
1
24

Figure 7

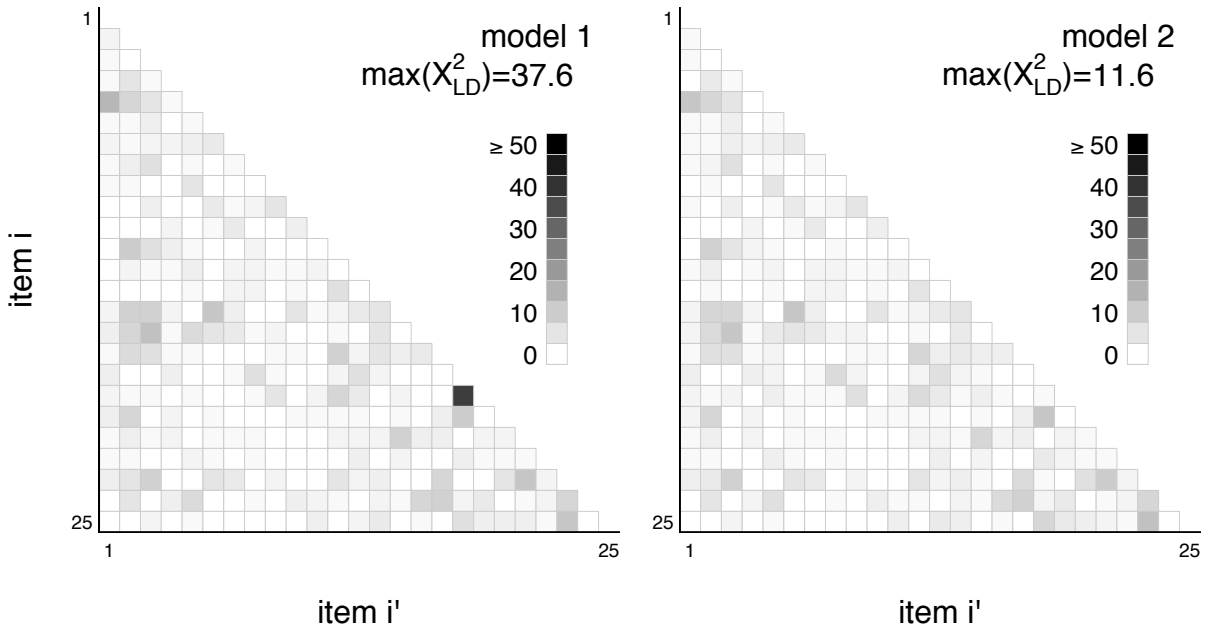
model 1
max(X^2_{LD})=37.6
model 2
max(X^2_{LD})=11.6
≥ 50
40
30
20
10
0
item i
item i'
1
25

Figure 8