Article

Journal of Educational and Behavioral Statistics
2016, Vol. 41, No. 1, pp. 3–26
DOI: 10.3102/1076998615621293

An Application of M_2 Statistic to Evaluate the Fit of Cognitive Diagnostic Models

Yanlou Liu
Beijing Normal University

Wei Tian
Beijing Normal University

Tao Xin
Beijing Normal University

The fit of cognitive diagnostic models (CDMs) to response data needs to be evaluated, since CDMs might yield misleading results when they do not fit the data well. Limited-information statistic M_2 *and the associated root mean square error of approximation (RMSEA*$_2$*) in item factor analysis were extended to evaluate the fit of CDMs. The findings suggested that the* M_2 *statistic has proper empirical Type I error rates and good statistical power, and it could be used as a general statistical tool. More importantly, we found that there was a strong linear relationship between mean marginal misclassification rates and RMSEA*$_2$ *when there was model–data misfit. The evidence demonstrated that .030 and .045 could be reasonable thresholds for excellent and good fit, respectively, under the saturated log-linear cognitive diagnosis model.*

Keywords: *cognitive diagnosis models; limited-information test statistics; goodness-of-fit; approximate fit*

Introduction

Cognitive diagnostic models (CDMs) are developed to provide fine-grained skills information for well-directed interventions. In fact, the chosen CDM, as a statistical model, needs to be evaluated regarding its validity with absolute model–data fit testing, in case the results are misleading when the model does not fit the data well (Rupp, Templin, & Henson, 2010). Generally speaking, the existing absolute model–data fit tests can be divided into item and test level. While item-level statistics have been the focus in recent work (e.g., Chen, de la Torre, & Zhang, 2013; de la Torre & Lee, 2013; Kunina-Habenicht, Rupp, & Wilhelm, 2012), test-level goodness-of-fit evaluation has been relatively

stagnant due to the fact that the full-information statistics χ^2 and G^2 are virtually useless in practice (e.g., Maydeu-Olivares, 2013) and other suitable alternatives remain underdeveloped.

The Pearson χ^2 test statistic and the likelihood ratio test statistic G^2 as full-information statistics are routinely used in the commercial software Mplus 7 (Muthén & Muthén, 2012). Both of them are computed from all possible response patterns (full contingence table). χ^2 and G^2 can be effective when all expected frequencies are large (the usual rule of thumb is that the expected frequency should exceed 5 in each cell). However, when the number of items is large or the number of respondents is small, so that the contingency table for possible responses is sparse, then χ^2 and G^2 cannot be trusted to test for lack of fit because the empirical Type I error rates are often different from the expected values under their reference asymptotic distributions (e.g., Maydeu-Olivares & Joe, 2005). Although alternatives to χ^2 and G^2 have been proposed to assess test-level fit (Rupp et al., 2010), much work remains to be done. For example, the Monte Carlo resampling technique (Templin & Henson, 2006) needs to obtain a stable empirical p value with multiple simulated data sets, which leads to an excessive amount of time. And the Bayesian posterior predictive model checking method (Sinharay, 2006; Sinharay & Almond, 2007) was found to be conservative and computationally intensive.

In practice, the limited-information method (Reiser, 1996; Reiser & Lin, 1999) is practicable since it uses the summary characteristics of the full contingence table (Cai & Hansen, 2013; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005). The limited-information statistics use only low-order marginal information in the contingency table to evaluate the model–data fit. The up-to-order r marginal information is used to develop the limited-information statistics M_r. The results showed that the M_2 statistic using the univariate and bivariate margins is enough for routine applications (Cai et al., 2006; Maydeu-Olivares & Joe, 2005). The model frequently does not exactly fit the data in practice; therefore, it is important to assess how well the model reflects reality (Brown, 2006; Hu & Bentler, 1998). Further, Maydeu-Olivares and Joe (2014) proposed up-to-order r root mean square error of approximation (RMSEA) fit indices RMSEA_r to assess the approximate goodness of fit. RMSEA_r can be estimated by using M_r, and Maydeu-Olivares and Joe (2014) recommended using RMSEA_2 to assess the approximate goodness of fit of the models for routine applications. Jurich (2014) conducted a small-scale simulation study to examine the statistical properties of M_2 under the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009); the results showed that M_2 had good Type I error control under the null conditions and high power to detect model misspecifications. However, no studies have systematically examined the performance of M_2 and RMSEA_2 in assessing model–data fit and degrees of model misfit, under CDMs.

This article aims to extend M_2 and RMSEA_2 in item response theory to those of CDM by taking the LCDM as an example. Firstly, we introduce LCDM and its marginal likelihood as the basis of computing CDMs' goodness-of-fit statistics. Secondly, we describe the idea of full- and limited-information model–data fit and approximate goodness-of-fit index. Thirdly, we present the results of simulation studies conducted to systematically evaluate the statistical properties of M_2 and the performance of RMSEA_2. Finally, as an empirical illustration, we present an analysis of fraction subtraction data originally reported by Tatsuoka (1990). This data set has been a commonly used example in cognitive diagnosis literature (e.g., DeCarlo, 2010; de la Torre, 2009, 2011; de la Torre & Douglas, 2008; Henson et al., 2009).

LCDM and Its Marginal Distribution

Suppose that a CDM test of J dichotomous scored items is administrated to N respondents, in which K binary attributes are diagnosed. Let the ith examinee's response vector be denoted by Bernoulli variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{iJ})'$ with realizations $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ})'$. A $J \times K$ Q-matrix is a binary specification of which attributes are required to solve each item. $\mathbf{q}_j = (q_{j1}, \ldots, q_{jk}, \ldots, q_{jK})'$ is the jth row of the Q-matrix, the element q_{jk} indicates the involvement relationship of the kth attribute for item j. The element $q_{jk} = 1$ means that attribute k is required to solve item j, and $q_{jk} = 0$ means that attribute k is not required to solve item j. The item parameters with which the saturated LCDM is associated can be denoted as $\boldsymbol{\beta}_j = (\lambda_{j,0}, \boldsymbol{\lambda}_j^{'})'$, where $\lambda_{j,0}$ is the intercept parameter, and $\boldsymbol{\lambda}_j$ are the main and interaction effect parameters for item j.

Statistically, the LCDM is essentially an item-based conditional probability of a correct response of the ith examinee for item j

$$P_j(\boldsymbol{\alpha}_i) = P(X_{ij} = 1|\boldsymbol{\alpha}_i) = \frac{\exp[\lambda_{j,0} + \boldsymbol{\lambda}_j^{'}\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}{1 + \exp[\lambda_{j,0} + \boldsymbol{\lambda}_j^{'}\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j)]}, \tag{1}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iK})'$ is the attribute mastery pattern of examinee i. The mapping function $\mathbf{h}$ is used to specify the linear combination of $\boldsymbol{\alpha}_i$ and q_j:

$$\begin{aligned}\boldsymbol{\lambda}_j^{'}\mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j) = &\sum_{k=1}^{K}\lambda_{j,1,(k)}\alpha_{ik}q_{jk} + \sum_{k=1}^{K-1}\sum_{k'=k+1}^{K}\lambda_{j,2,(k,k')}\alpha_{ik}\alpha_{ik'}q_{jk}q_{jk'} + \ldots \\ &+ \lambda_{j,K_j,(1,\ldots,K_j)}\prod_{k=1}^{K_j}\alpha_{ik}q_{jk},\end{aligned} \tag{2}$$

where $K_j = \sum_{k=1}^{K} q_{jk}$ is the number of required attributes of the j th item defined by the Q-matrix. In Equation 2, the subscript following the first comma represents the effect level; the subscript(s) in the parentheses represents the included

attribute in the effect. For example, if $\mathbf{q}_j = (0,1,0,1)^{'}$, $\boldsymbol{\alpha}_i = (1,1,0,0)^{'}$, that is to say the second and the fourth attributes are required by item j and the ith examinee has mastered the first-two attributes, but the first attribute is not required by the item, then $\boldsymbol{\lambda}_j^{'} = \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j) = \lambda_{j,1,(2)}$, which is the main effect for item j and Attribute 2. However, if $\boldsymbol{\alpha}_i = (0,1,0,1)^{'}$, which indicates examinee i has mastery all the required attributes, then, $\boldsymbol{\lambda}_j^{'} = \mathbf{h}(\boldsymbol{\alpha}_i, \mathbf{q}_j) = \lambda_{j,1,(2)} + \lambda_{j,1,(4)} + \lambda_{j,2,(2,4)}$, for the unconstraint LCDM, $\lambda_{j,1,(2)}$ and $\lambda_{j,1,(4)}$ are the main effects associated with Attributes 2 and 4, respectively, $\lambda_{j,2,(2,4)}$ is the two-way interaction effect associated with Attributes 2 and 4.

Further, assuming local independence, the likelihood function for examinee i, conditional on the attribute mastery pattern $\boldsymbol{\alpha}_i$ is:

$$P(\mathbf{X}_i|\boldsymbol{\alpha}_i) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_i)^{x_{ij}}[1 - P_j(\boldsymbol{\alpha}_i)]^{1-x_{ij}}. \tag{3}$$

The marginal probability of the ith examinee's response vector can be written as:

$$\pi_{\mathbf{X}_i}(\boldsymbol{\gamma}) = P(\mathbf{X}_i) = \sum_{l=1}^{L} P(\mathbf{X}_i|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l). \tag{4}$$

In the above expression, $p(\boldsymbol{\alpha}_l)$ is the probability of attribute mastery pattern $\boldsymbol{\alpha}_l$, $L = 2^K$ is the total number of possible attribute mastery patterns, and $\boldsymbol{\gamma} = (\boldsymbol{\beta}_1^{'}, \ldots, \boldsymbol{\beta}_J^{'}, \boldsymbol{\eta}^{'})^{'}$ is the vector of the model parameters. There is a constraint on the attribute mastery pattern probabilities,

$$\sum_{l=1}^{L} p(\boldsymbol{\alpha}_l) = 1. \tag{5}$$

To accommodate this sum-to-one constraint, the following expression is used:

$$p(\boldsymbol{\alpha}_l) = \frac{\exp(\eta_l)}{\sum_{l=1}^{L} \exp(\eta_l)}, \tag{6}$$

and one identifying constraint has to be imposed on the model parameter η_l, for example, the value for η_L is fixed at 0. The model parameter estimates of LCDM can be estimated by using marginal maximum likelihood estimation; for saturated LCDM, the number of free parameters is:

$$F = \sum_{j=1}^{J} 2^{K_j} + (L - 1). \tag{7}$$

The model-predicted probability $\pi_{\mathbf{X}_i}(\hat{\boldsymbol{\gamma}})$ can be obtained by substituting model parameter estimates into Equation 4.

The Limited Information as an Alternative of Full Information

Full-Information Statistics

The exact model fit is used to determine whether CDMs can precisely predict the phenomena provided by the data (Sinharay & Almond, 2007). In other words, the model-fit statistics compare the following model-predicted probabilities and the observed proportions associated with the contingency table:

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_h \\ \vdots \\ \pi_H \end{pmatrix}, \ \boldsymbol{\pi}(\hat{\boldsymbol{\gamma}}) = \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \vdots \\ \hat{\pi}_h \\ \vdots \\ \hat{\pi}_H \end{pmatrix}, \ \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_h \\ \vdots \\ p_H \end{pmatrix}.$$

Here, $\boldsymbol{\pi}$ is the vector of cell probabilities, $\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$ is the maximum likelihood estimator of $\boldsymbol{\pi}$, $\mathbf{p}$ is the vector of sample cell proportions, and $H = 2^J$ is the number of possible response patterns. For simplicity, let $\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{\pi}}$ be used in subsequent expressions with no resulting confusion.

The aim of the exact model-fit statistics is to test the simple or composite null hypothesis. We focused on the composite context because it is encountered more frequently in psychological and educational assessments. The composite null hypothesis of the full information is:

$$\begin{aligned} H_0 &: \boldsymbol{\pi} = \boldsymbol{\pi}(\boldsymbol{\gamma}) \text{ for some } \boldsymbol{\gamma}, \\ H_1 &: \boldsymbol{\pi} \neq \boldsymbol{\pi}(\boldsymbol{\gamma}) \text{ for any } \boldsymbol{\gamma}, \end{aligned}$$

where $\boldsymbol{\gamma}$ is the vector of model parameters determined in advance. In practice, $\boldsymbol{\gamma}$ is typically estimated from the data. The full-information statistics:

$$\chi^2 = N\sum_{h=1}^{H} \frac{(p_h - \hat{\pi}_h)^2}{\hat{\pi}_h} \text{ and } G^2 = 2N\sum_{h=1}^{H} p_h \ln\left(\frac{p_h}{\hat{\pi}_h}\right),$$

compare the contingency table using full information from all response patterns; degrees of freedom of χ^2 and G^2 are $H - F - 1$. It should be noted that χ^2 and G^2 should not be used if some expected cell counts are small.

According to Maydeu-Olivares and Joe (2014), once the model parameters $\boldsymbol{\gamma}$ have been estimated by ML, the full-information RMSEA_J can be obtained by using χ^2 and its corresponding *df*:

$$\text{RMSEA}_J = \sqrt{\text{Max}\left(\frac{\hat{\chi}^2 - df}{N \times df}, 0\right)}. \tag{8}$$

A 90% confidence interval (CI) for RMSEA_J is (Browne & Cudeck, 1993):

$$\left(\sqrt{\frac{\hat{\mathcal{L}}}{N \times df}}; \sqrt{\frac{\hat{\mathcal{U}}}{N \times df}}\right), \tag{9}$$

where $\hat{\mathcal{L}}$ and $\hat{\mathcal{U}}$ are the roots of the noncentral χ^2 distribution functions:

$$\mathcal{F}_{\chi^2}(\hat{\chi}^2; df, \hat{\mathcal{L}}) = .95 \text{ and } \mathcal{F}_{\chi^2}(\hat{\chi}^2; df, \hat{\mathcal{U}}) = .05. \tag{10}$$

Limited-Information Statistics

By contrast, limited-information statistics were developed as an alternative to address the frequently encountered sparseness phenomenon by compressing the contingency table. Specifically, the H cells of marginal probabilities associated with the full response patterns are significantly reduced to smaller ones. Essentially, the limited-information statistics are used as a statistical concept of lower-order margins.

Symbolically, let the first-order marginal probabilities $\dot{\boldsymbol{\pi}}_1$ represent the marginal probabilities of correctly answering each single item in the test. The second-order marginal probabilities $\dot{\boldsymbol{\pi}}_2$ represent the marginal probabilities of correctly answering each item pair. The rth-order marginal probabilities $\dot{\boldsymbol{\pi}}_r$ represent the marginal probabilities of correctly answering r items simultaneously. One easy way to compute the lower marginal probabilities $\dot{\boldsymbol{\pi}} = (\dot{\boldsymbol{\pi}}_1^{'}, \dot{\boldsymbol{\pi}}_2^{'}, \ldots, \dot{\boldsymbol{\pi}}_r^{'}, \ldots, \dot{\boldsymbol{\pi}}_J^{'})^{'}$ from the original contingency table is to construct a $(2^J - 1) \times 2^J$ dimensional indicator matrix $\mathbf{T} = (\dot{\mathbf{T}}_1^{'}, \dot{\mathbf{T}}_2^{'}, \ldots, \dot{\mathbf{T}}_r^{'}, \ldots, \dot{\mathbf{T}}_J^{'})^{'}$ with elements of zeroes or ones (Cai et al., 2006; Maydeu-Olivares & Joe, 2005; Reiser, 1996), that is to say, $\dot{\boldsymbol{\pi}} = \mathbf{T}\boldsymbol{\pi}$. $\dot{\boldsymbol{\pi}}_1 = (\dot{\pi}_1, \ldots, \dot{\pi}_j, \ldots, \dot{\pi}_J)^{'}$ represents the $\binom{J}{1}$-dimensional first-order marginal probabilities.

$$\dot{\pi}_j = P(X_j = 1) \tag{11}$$

is the marginal probability of correctly answering the jth item. $\dot{\boldsymbol{\pi}}_2$ is the vector of $\binom{J}{2}$-dimensional second-order marginal probabilities. The element:

$$\dot{\pi}_{a,b} = P(X_a = 1, X_b = 1), 1 \leq a < b \leq J, \tag{12}$$

contained in $\dot{\boldsymbol{\pi}}_2$ is the marginal probability of correctly answering the ath item and bth item simultaneously. $\dot{\boldsymbol{\pi}}_r$ stands for the $\binom{J}{r}$-dimensional rth-order marginal probabilities. Let vector $\boldsymbol{\pi}_r = (\dot{\boldsymbol{\pi}}_1^{'}, \ldots, \dot{\boldsymbol{\pi}}_r^{'})^{'}$ denotes up-to-order r $(r \leq J)$

joint marginal probabilities, then $\boldsymbol{\pi}_r = \mathbf{T}_r\boldsymbol{\pi}$, where $\mathbf{T}_r = (\dot{\mathbf{T}}_1^{'}, \ldots, \dot{\mathbf{T}}_r^{'})^{'}$ is a $R \times 2^J$ matrix, $R = \sum_{r=1}^{r} \binom{J}{r}$.

For example, if 3 items were designed for a diagnostic test, then the lower-order marginal probabilities would be:

$$\dot{\boldsymbol{\pi}} = \begin{pmatrix} \dot{\boldsymbol{\pi}}_1 \\ \dot{\boldsymbol{\pi}}_2 \\ \dot{\boldsymbol{\pi}}_3 \end{pmatrix} = \begin{pmatrix} \dot{\pi}_1 \\ \dot{\pi}_2 \\ \dot{\pi}_3 \\ \dot{\pi}_{1,2} \\ \dot{\pi}_{1,3} \\ \dot{\pi}_{2,3} \\ \dot{\pi}_{1,2,3} \end{pmatrix} = \mathbf{T}\boldsymbol{\pi} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \pi_{(0,0,0)} \\ \pi_{(1,0,0)} \\ \pi_{(0,1,0)} \\ \pi_{(0,0,1)} \\ \pi_{(1,1,0)} \\ \pi_{(1,0,1)} \\ \pi_{(0,1,1)} \\ \pi_{(1,1,1)} \end{pmatrix}. \quad (13)$$

When the model parameters are estimated from the data under the necessary regularity assumptions on the model, the ML estimator $\hat{\boldsymbol{\gamma}}$ is consistent and is asymptotically normally distributed (Bishop, Fienberg, & Holland, 2007):

$$\sqrt{N}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{I}^{-1}), \quad (14)$$

$\mathcal{I} = \boldsymbol{\Delta}^{'}\{\text{diag}[\boldsymbol{\pi}(\boldsymbol{\gamma})]\}^{-1}\boldsymbol{\Delta}$ is the Fisher information matrix, $\text{diag}[\boldsymbol{\pi}(\boldsymbol{\gamma})]$ is a diagonal matrix of $\boldsymbol{\pi}(\boldsymbol{\gamma})$, $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}^{'}$ is the Jacobian matrix, and $\xrightarrow{d}$ denotes convergence in distribution. The limited-information goodness-of-fit statistics compare the merged model-predicted marginal probabilities and the corresponding sample-observed counterparts. The vector of up-to-order r residuals $\hat{\mathbf{e}}_r = \mathbf{T}_r(\mathbf{p} - \hat{\boldsymbol{\pi}})$ is a linear transformation of the cell residual vector, which is also asymptotically normally distributed (Maydeu-Olivares & Joe, 2005; Reiser, 1996),

$$\sqrt{N}(\hat{\mathbf{e}}_r) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_r), \quad (15)$$

where $\boldsymbol{\Sigma}_r = \mathbf{T}_r(\boldsymbol{\Gamma} - \boldsymbol{\Delta}\mathcal{I}^{-1}\boldsymbol{\Delta}^{'})\,\mathbf{T}_r^{'} = \boldsymbol{\Gamma}_r - \boldsymbol{\Delta}_r\mathcal{I}^{-1}\boldsymbol{\Delta}_r^{'}$, $\boldsymbol{\Gamma} = \text{diag}[\boldsymbol{\pi}(\boldsymbol{\gamma})] - \boldsymbol{\pi}(\boldsymbol{\gamma})\boldsymbol{\pi}(\boldsymbol{\gamma})^{'}$ is the covariance matrix of $\boldsymbol{\pi}(\boldsymbol{\gamma})$, and $\boldsymbol{\Delta}_r = \mathbf{T}_r\boldsymbol{\Delta}$.

Limited-information statistics M_r based on up-to-order r marginal residuals can be written as follows (Maydeu-Olivares & Joe, 2005):

$$M_r = N(\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r)^{'}\hat{\mathbf{C}}_r(\mathbf{p}_r - \hat{\boldsymbol{\pi}}_r), \quad (16)$$

where $\hat{\mathbf{C}}_r = \hat{\boldsymbol{\Delta}}_r^{(c)}(\hat{\boldsymbol{\Delta}}_r^{(c)'}\hat{\boldsymbol{\Gamma}}_r\hat{\boldsymbol{\Delta}}_r^{(c)})^{-1}\hat{\boldsymbol{\Delta}}_r^{(c)'}$, $\hat{\boldsymbol{\Delta}}_r^{(c)}$ is an $R \times (R - F)$ orthogonal complement to $\hat{\boldsymbol{\Delta}}_r$ (Browne, 1984). Under H_0, $M_r \xrightarrow{d} \chi^2_{R-F}$, and the corresponding df is $R - F$ (Khatri, 1966).

Evaluate the Fit of CDMs Using M_2 Statistic

In CDMs, it is often the case that both the item parameters and the probabilities of attribute mastery pattern are estimated simultaneously from the data using

ML procedure. To test the composite null hypotheses, the limited-information statistic M_2 based upon up-to-order two residuals has been recommended by researchers (Cai & Hansen, 2013; Maydeu-Olivares & Joe, 2005). Therefore, the composite hypothesis is:

$$H_0 : \boldsymbol{\pi}_2 = \boldsymbol{\pi}_2(\boldsymbol{\gamma}) \text{ for some } \boldsymbol{\gamma},$$
$$H_1 : \boldsymbol{\pi}_2 \neq \boldsymbol{\pi}_2(\boldsymbol{\gamma}) \text{ for any } \boldsymbol{\gamma}.$$

After the model parameters have been estimated from the data, the model-predicted marginal probabilities of up-to-order two response patterns are:

$$\hat{\boldsymbol{\pi}}_2 = \begin{pmatrix} \hat{\dot{\boldsymbol{\pi}}}_1 \\ \hat{\dot{\boldsymbol{\pi}}}_2 \end{pmatrix}. \tag{17}$$

The observed proportions associated with the univariate and bivariate model-predicted marginal probabilities are:

$$\mathbf{p}_2 = \begin{pmatrix} \dot{\mathbf{p}}_1 \\ \dot{\mathbf{p}}_2 \end{pmatrix}. \tag{18}$$

The asymptotic covariance matrix $\hat{\boldsymbol{\Gamma}}_2$ of $\sqrt{N}(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)$ can be partitioned into four blocks, $\hat{\boldsymbol{\Gamma}}_{(1,1)} = \sqrt{N}\text{Acov}(\dot{\mathbf{p}}_1)$, $\hat{\boldsymbol{\Gamma}}_{(2,1)} = \hat{\boldsymbol{\Gamma}}'_{(1,2)} = \sqrt{N}\text{Acov}(\dot{\mathbf{p}}_2, \dot{\mathbf{p}}_1)$, and $\hat{\boldsymbol{\Gamma}}_{(2,2)} = \sqrt{N}\text{Acov}(\dot{\mathbf{p}}_2, \dot{\mathbf{p}}_2)$ according to the partitioning of $\hat{\boldsymbol{\pi}}_2$:

$$\hat{\boldsymbol{\Gamma}}_2 = \begin{pmatrix} \hat{\boldsymbol{\Gamma}}_{(1,1)} & \hat{\boldsymbol{\Gamma}}_{(1,2)} \\ \hat{\boldsymbol{\Gamma}}_{(2,1)} & \hat{\boldsymbol{\Gamma}}_{(2,2)} \end{pmatrix}, \tag{19}$$

where Acov( ) denotes an asymptotic covariance matrix. The elements in the submatrices $\hat{\boldsymbol{\Gamma}}_{(1,1)}$, $\hat{\boldsymbol{\Gamma}}_{(2,1)}$, and $\hat{\boldsymbol{\Gamma}}_{(2,2)}$ are, respectively:

$$\sqrt{N}\text{Acov}(\dot{p}_a, \dot{p}_b) = \hat{\dot{\pi}}_{a,b} - \hat{\dot{\pi}}_a \hat{\dot{\pi}}_b, \tag{20}$$
$$\sqrt{N}\text{Acov}(\dot{p}_{a,b}, \dot{p}_c) = \hat{\dot{\pi}}_{a,b,c} - \hat{\dot{\pi}}_{a,b} \hat{\dot{\pi}}_c, \ a < b, \tag{21}$$
$$\sqrt{N}\text{Acov}(\dot{p}_{a,b}, \dot{p}_{c,d}) = \hat{\dot{\pi}}_{a,b,c,d} - \hat{\dot{\pi}}_{a,b} \hat{\dot{\pi}}_{c,d}, \ a < b, \ c < d, \tag{22}$$

and

$$\hat{\dot{\pi}}_a = P(X_a = 1), \ \hat{\dot{\pi}}_b = P(X_b = 1) \text{ and } \hat{\dot{\pi}}_c = P(X_c = 1), \tag{23}$$

are the marginal probabilities of correctly answering the a th, b th, and c th item, respectively:

$$\hat{\dot{\pi}}_{a,b} = P(X_a = 1, X_b = 1) \text{ and } \hat{\dot{\pi}}_{c,d} = P(X_c = 1, X_d = 1), \tag{24}$$

are the marginal probabilities of correctly answering the ath and bth item pair and the cth and dth item pair, respectively;

$$\hat{\dot{\pi}}_{a,b,c} = P(X_a = 1, X_b = 1, X_c = 1), \tag{25}$$

is the marginal probability of correctly answering the ath item, bth item, and cth item simultaneously:

$$\hat{\pi}_{a,b,c,d} = P(X_a = 1, X_b = 1, X_c = 1, X_d = 1), \tag{26}$$

is the marginal probability of correctly answering the ath item, bth item, and cth item, and dth item simultaneously.

The $R \times F$ Jacobian matrix $\hat{\boldsymbol{\Delta}}_2$ is the first-order partial derivatives of $\hat{\boldsymbol{\pi}}_2$,

$$\hat{\boldsymbol{\Delta}}_2 = \frac{\partial \hat{\boldsymbol{\pi}}_2}{\partial \hat{\boldsymbol{\gamma}}'}, \tag{27}$$

where $R = \sum_{r=1}^{2} \binom{J}{r}$. The Jacobian matrix $\hat{\boldsymbol{\Delta}}_2$ is of full column rank F. According to the partitioning of $\hat{\boldsymbol{\pi}}_2$, $\hat{\boldsymbol{\Delta}}_2$ can also be partitioned into four blocks:

$$\hat{\boldsymbol{\Delta}}_2 = \begin{pmatrix} \hat{\boldsymbol{\Delta}}_{(1,1)} & \hat{\boldsymbol{\Delta}}_{(1,2)} \\ \hat{\boldsymbol{\Delta}}_{(2,1)} & \hat{\boldsymbol{\Delta}}_{(2,2)} \end{pmatrix} = \begin{pmatrix} \frac{\partial \hat{\boldsymbol{\pi}}_1}{\partial \hat{\boldsymbol{\beta}}'} & \frac{\partial \hat{\boldsymbol{\pi}}_1}{\partial \hat{\boldsymbol{\eta}}'} \\ \frac{\partial \hat{\boldsymbol{\pi}}_2}{\partial \hat{\boldsymbol{\beta}}'} & \frac{\partial \hat{\boldsymbol{\pi}}_2}{\partial \hat{\boldsymbol{\eta}}'} \end{pmatrix} \tag{28}$$

The generic entries of $\hat{\boldsymbol{\Delta}}_{(1,1)}$ and $\hat{\boldsymbol{\Delta}}_{(1,2)}$ of the first-order marginal probability are:

$$\frac{\partial \hat{\pi}_a}{\partial \hat{\boldsymbol{\beta}}_a'} = \sum_{l=1}^{L} P(X_a|\boldsymbol{\alpha}_l)[1 - P(X_a|\boldsymbol{\alpha}_l)]p(\boldsymbol{\alpha}_l) \tag{29}$$

and

$$\frac{\partial \hat{\pi}_a}{\partial \widehat{\eta}_l} = \sum_{l'=1}^{L} P(X_a|\boldsymbol{\alpha}_{l'}) \frac{\partial p(\boldsymbol{\alpha}_{l'})}{\partial \widehat{\eta}_l}. \tag{30}$$

The elements of $\hat{\boldsymbol{\Delta}}_{(2,1)}$ and $\hat{\boldsymbol{\Delta}}_{(2,2)}$ of the Jacobian matrix for the second-order are:

$$\frac{\partial \hat{\pi}_{a,b}}{\partial \hat{\boldsymbol{\beta}}_a'} = \sum_{l=1}^{L} P(X_b|\boldsymbol{\alpha}_l)P(X_a|\boldsymbol{\alpha}_l)[1 - P(X_a|\boldsymbol{\alpha}_l)]p(\boldsymbol{\alpha}_l), \tag{31}$$

$$\frac{\partial \hat{\pi}_{a,b}}{\partial \hat{\boldsymbol{\beta}}_b'} = \sum_{l=1}^{L} P(X_a|\boldsymbol{\alpha}_l)P(X_b|\boldsymbol{\alpha}_l)[1 - P(X_b|\boldsymbol{\alpha}_l)]p(\boldsymbol{\alpha}_l), \tag{32}$$

and

$$\frac{\partial \hat{\pi}_{a,b}}{\partial \widehat{\eta}_l} = \sum_{l'=1}^{L} P(X_a|\boldsymbol{\alpha}_{l'})P(X_b|\boldsymbol{\alpha}_{l'}) \frac{\partial p(\boldsymbol{\alpha}_{l'})}{\partial \widehat{\eta}_l}. \tag{33}$$

TABLE 1.
True Item Parameters

K_j	$\lambda_{j,0}$	$\lambda_{j,1,(k)}$	$\lambda_{j,2,(k,k')}$	$\lambda_{j,3,(k,k',K_j)}$
1	−1.39	3.58		
2	−1.39	1.39	.80	
3	−1.39	.51	.51	.52

Note. K_j is the number of the required attributes by the jth item. $\lambda_{j,0}$ is the intercept parameter. $\lambda_{j,1,(k)}$ is the main effect parameter. $\lambda_{j,2,(k,k')}$ is the two-way interaction parameter, $k \neq k'$. $\lambda_{j,3,(k,k'K_j)}$ is the three-way interaction parameter.

According to Equation 16, M_2 for CDMs can be constructed:

$$M_2 = N(\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2)' \hat{\mathbf{C}}_2 (\mathbf{p}_2 - \hat{\boldsymbol{\pi}}_2), \tag{34}$$

where $\hat{\mathbf{C}}_2 = \hat{\boldsymbol{\Delta}}_2^{(c)} (\hat{\boldsymbol{\Delta}}_2^{(c)'} \hat{\boldsymbol{\Gamma}}_2 \hat{\boldsymbol{\Delta}}_2^{(c)})^{-1} \hat{\boldsymbol{\Delta}}_2^{(c)'}$. RMSEA_2 can be obtained using the observed $\hat{M}_2$ and the corresponding df:

$$\text{RMSEA}_2 = \sqrt{\text{Max}\left(\frac{\hat{M}_2 - df}{N \times df}, 0\right)}. \tag{35}$$

And a 90% CI for RMSEA_2 is:

$$\left(\sqrt{\frac{\hat{\mathcal{L}}}{N \times df}}; \sqrt{\frac{\hat{\mathcal{U}}}{N \times df}}\right); \tag{36}$$

here, $\hat{\mathcal{L}}$ and $\hat{\mathcal{U}}$ are the roots of the functions:

$$\mathcal{F}_{\chi^2}(\hat{M}_2; df, \hat{\mathcal{L}}) = .95 \text{ and } \mathcal{F}_{\chi^2}(\hat{M}_2; df, \hat{\mathcal{U}}) = .05. \tag{37}$$

Simulation Studies

A series of Monte Carlo simulations were conducted using *R* software (R Core Team, 2014). R code for computing M_2 and RMSEA_2 of the LCDM framework and generalized deterministic inputs, and the noisy "and" gate model (de la Torre, 2011) framework are all available upon request from the first author. For each simulation, two sample sizes, $N = 1{,}000$ and $N = 5{,}000$, were considered. There were 500 replications in each condition. The data were all generated using the saturated LCDM model. Table 1 summarizes the true item parameters. Table 2 presents the summary of the true probabilities of responding correctly to items that required one, two, or three attributes for examinees that had mastered zero, one, two, and three attributes.

The purpose of the first study was to investigate the performance of χ^2, G^2, and M_2 under the null condition of correct model specifications. Simulation 1 was

TABLE 2.
True Item Response Probabilities

	Number of Attributes Mastered by Examinee			
K_j	0	1	2	3
1	.20	.90		
2	.20	.50	.90	
3	.20	.29	.53	.90

Note. K_j is the number of required attributes by the jth item.

conducted to examine the Type I error rates of χ^2, G^2, and M_2, when the test lengths were relatively small and to examine the Type I error rates of M_2, when the test lengths were relatively large. In the second study, simulation 2 was conducted to illustrate the power of M_2, when the Q-matrix of the model was misspecified. The purpose of the third study was to explore the performance of RMSEA_2 under more comprehensive Q-matrix misspecification conditions. Simulation 3 was conducted to examine the performance of RMSEA_2 systematically.

Type I Error Rates

In this section, a simulation was performed to examine the empirical Type I error rates of χ^2, G^2, and M_2. We compared the Type I error rates of the three statistics under nonsparse and sparse conditions. In order to test the robustness of the M_2 statistic, a multidimensional normal distribution was specified for the attributes; the mean vectors were randomly chosen from the uniform distribution $\mu(-.5,\ .5)$, the correlation coefficients between the two attributes were randomly chosen from $\mu(.5,\ .8)$, which are typical correlation coefficients between attributes (Kunina-Habenicht et al., 2012; Sinharay, Puhan, & Haberman, 2011), and the attribute vectors were dichotomized at 0.

Simulation conditions. For this simulation, three factors were manipulated: Four Test Lengths $\times$ Two Sample Sizes $\times$ Three Statistics. (a) Test length: 6, 15, 30, and 50 items; (b) sample size: 1,000 and 5,000 examinees; (c) statistic: χ^2, G^2, and M_2. When $J = 30$, the number of cells in the contingency table equaled more than 1 billion, so χ^2 and G^2 were not considered in conditions with $J = 30$ and $J = 50$. There were a total of 16 data generating conditions that were considered.

For conditions of 6 items, there were two latent attributes. The Q-matrix for 6 items is presented in Table 3. For the 15, 30, and 50 item conditions, there were five attributes, and the Q-matrices are presented in Appendix Table A1. For each Q-matrix, the number of items for measuring each attribute was equal, and three attributes were required by items at most.

TABLE 3.
Q-Matrix for J = 6, K = 2

Item	α_1	α_2
1	1	0
2	0	1
3	1	1
4	1	0
5	0	1
6	1	1

TABLE 4.
*Type I Error Rates for Nonsparse Contingency Tables (*J = 6, K = 2)

					Empirical Rejection Rate				
N	Statistic	*df*	Mean	Var/2	.010	.050	.100	.200	.250
1,000	M_2	2	1.99	2.33	.010	.054	.100	.188	.244
	χ^2	44	44.02	44.85	.004	.050	.112	.218	.262
	G^2	44	46.39	50.16	.014	.088	.164	.308	.356
5,000	M_2	2	1.96	2.13	.014	.050	.106	.192	.239
	χ^2	44	44.28	42.66	.012	.052	.110	.200	.248
	G^2	44	44.58	43.32	.014	.060	.112	.216	.250

Note. N = the sample size; *df* = the degrees of freedom; Var/2 = variance/2.

Results. Since we assumed that the attributes were moderately or highly correlated, some of the expected frequencies of the response patterns were certainly smaller than N/H. Table 4 shows the results of the observed Type I error rates when $J = 6$ for χ^2, G^2, and M_2 at five significance levels: .010, .050, .100, .200, and .250. When the test length was six and the sample size was 1,000, although the degree of sparseness was $N/H = 1,000/2^6 = 15.625$, some of the expected frequencies in the contingency table might be smaller than N/H, and the performance of χ^2 was better than G^2 (Koehler & Larntz, 1980). When the sample size was 5,000, the distributions of χ^2 and G^2 were close to the asymptotic χ^2 distributions, so that the two statistics had good Type I error rate control under this nonsparse contingency table condition. Table 4 also illustrates that when $J = 6$, the performance of M_2 was comparable or superior to χ^2 and G^2.

When 15 items were considered in a test, the number of all response patterns was $H = 2^{15}$; that is to say even though the sample size was 5,000, the degree of sparseness N/H was .153. When the contingency table was sparse, the Type I error rates under the five significant levels are shown in Table 5. It is evident that the severe sparseness led to the empirical Type I error rates of χ^2 and G^2 not matching their expected rates. Even

TABLE 5.
Type I Error Rates for Sparse Contingency Tables (J = 15)

					Empirical Rejection Rate				
N	Statistic	*df*	Mean	Var/2	.010	.050	.100	.200	.250
1,000	M_2	19	19.30	19.29	.010	.046	.108	.218	.260
	χ^2	32,681	32,569.81	1,922,140.15	.326	.358	.372	.398	.410
	G^2	32,681	4,488.06	11,048.10	.000	.000	.000	.000	.000
5,000	M_2	19	19.17	21.15	.016	.070	.108	.214	.266
	χ^2	32,681	32,783.85	332,217.40	.276	.346	.400	.448	.470
	G^2	32,681	11,944.11	56,238.43	.000	.000	.000	.000	.000

Note. n = the sample size; *df* = the degrees of freedom; Var/2 = variance/2.

TABLE 6.
Type I Error Rates for Sparse Contingency Tables (J = 30, 50)

					Empirical Rejection Rate				
J	*N*	*df*	Mean	Var/2	.010	.050	.100	.200	.250
30	1,000	294	295.80	279.40	.012	.042	.096	.210	.284
	5,000	294	294.52	299.49	.008	.052	.114	.220	.276
50	1,000	984	987.49	1,038.64	.014	.066	.122	.228	.28
	5,000	984	983.89	1,074.40	.012	.058	.116	.192	.252

Note. J = the test length; N = the sample size; *df* = the degrees of freedom; Var/2 = Variance/2.

though the test length was relatively small, χ^2 and G^2 did not have good Type I error rate control. However, the Type I error rates of M_2 were close to the nominal levels.

We also investigated the performance of M_2, under large test length conditions. As is shown in Table 6, when $J = 30$ and $J = 50$, the Type I error rates of M_2 were reasonably close to the nominal levels. Integrating Tables 4, 5, and 6, the empirical Type I error rates of M_2 were found to be stable and accurate.

Power to Detect Model Misspecification

Some studies have revealed that there are many sources of model–data misfit, such as CDM misspecification and Q-matrix misspecification (Chen et al., 2013; de la Torre & Lee, 2013; Kunina-Habenicht et al., 2012). Kunina-Habenicht, Rupp, and Wilhelm (2012) have shown that if the data were generated by the saturated model, the exclusion of the interaction effects did not have a noticeable impact on the marginal correct classification rates; however, the misspecification of Q-matrix had a significant effect on classification accuracy. In this section, in order to illustrate the performance of M_2, under the incorrect specification of the

TABLE 7.
The Random Balance Design of Q-Matrix Misspecification

K_j	Alterations	Note
1	$q_{jk} = 0 \rightarrow q_{jk} = 1$	Overspecification
2	$q_{jk} = 1 \rightarrow q_{jk} = 0$	Underspecification
3	$q_{jk} = 0 \rightarrow q_{jk} = 1, q_{jk'} = 1 \rightarrow q_{jk'} = 0$	Both

Note. K_j is the number of required attributes by the jth item. q_{jk} is the entry in the jth row and kth column of the Q-matrix. $q_{jk'}$ is the entry in the jth row and k' th column of the Q-matrix, $k \neq k'$.

Q-matrix conditions, a simulation was conducted to examine the power of M_2. In this simulation, the data were generated using LCDM and the Q-matrices are provided in Appendix Table A1.

Simulation conditions. The following four factors were manipulated in the simulation: Three Test Length Conditions × Two Q-matrix Misspecification Types × Two Attribute Correlation Conditions × Two Sample Sizes. (a) Test length: 15, 30, and 50 items; (b) Q-matrix misspecification type: the random design and the random balance design; (c) attribute correlation: .5 and .8; and (d) sample size: 1,000 and 5,000 examinees. A total of 24 data generating conditions were examined.

For Q-matrix misspecification, two levels were considered: The first was that 20% of the Q-matrix elements were randomly assigned by limiting the maximum number of required attributes to 3 and the minimum to 1, we will refer to this as a random design. In the second condition, 20% of the elements of the Q-matrix were misspecified; specifically, the overspecified items were chosen randomly from the items measuring only one attribute, and both under- and overspecified items were chosen randomly from items measuring three attributes, which we will refer to as a random balance design. The random balance design was driven by the Q-matrix misspecification in the previous studies (Chen et al., 2013), and the summary of the random balance design of Q-matrix misspecification is shown in Table 7. The misspecified Q-matrix was generated for each replication. Kunina-Habenicht et al. (2012) reported that in their mathematics and network engineering assessments, the correlations between attributes were all higher than .5 and, some even exceeded .8, and in their simulation studies with LCDM, the correlations between attributes were set to $\rho = .5$ and $\rho = .8$. In order to compare the performance of M_2 under different attribute correlation conditions, we followed the simulation design used by Kunina-Habenicht et al. (2012), and the strength of the attribute tetrachoric correlations were set to $\rho = .5$ and $\rho = .8$ to reflect the low and high correlations between attributes.

Results. The empirical rejection rates of M_2 were 1 under the random balance design, and the rejection rates were all above .990 when $J = 30$ and $J = 50$. Thus, only the empirical power results of M_2 for $J = 15$ under the random design are

TABLE 8.
The Empirical Rejection Rates of M_2 *for* J = *15 Under the Random Design*

		Empirical Rejection Rate				
ρ	*N*	.010	.050	.100	.200	.250
.5	1,000	.746	.818	.848	.874	.882
.5	5,000	.916	.936	.946	.966	.966
.8	1,000	.532	.668	.732	.79	.808
.8	5,000	.892	.926	.928	.94	.948

Note. ρ = the attribute correlation. *N* = the sample size.

presented in Table 8. As shown in Table 8, the statistical power for ρ = .5 was higher than ρ = .8, and with the increase of the sample size, the power of M_2 was also strengthened. In addition, the results revealed that M_2 was a powerful tool for detecting model–data misfit.

The Performance of RMSEA$_2$

Overall goodness-of-fit index provides us the information regarding the degree of discrepancy between observed and expected responses. However, it is realistic to expect that CDMs might not always fit the data exactly in practice. As illustrated in the section on power to detect model misspecification, the statistic M_2 has high power to detect model misspecification. Therefore, it is important to provide an effect-size estimator to evaluate whether model–data fit is good enough for practitioners or not. In order to systematically examine the performance of $RMSEA_2$ in LCDM, Simulation 3 was conducted with particular attention to the relationship between the $RMSEA_2$ and the mean marginal misclassification rates (MMMRs) under more comprehensive Q-matrix misspecification conditions. Specifically, three levels of the percentage of Q-matrix misspecification: 10% (small), 20% (medium), and 30% (large) and two types of Q-matrix misspecification: Random design and random balance design were manipulated.

Simulation conditions. The following five factors were manipulated in the simulations: (a) test length: 15, 30, and 50 items; (b) percentage of Q-matrix misspecification: 10%, 20%, and 30%; (c) Q-matrix misspecification type: the random design and the random balance design; (d) attribute correlation: .5 and .8; (e) sample size: 1,000 and 5,000 examinees. Since the required numbers of item were larger than the test length, the 30% random balance design condition was not included in Simulation 3. There were a total of 60 conditions.

Results. The primary focus was to examine the relationship between $RMSEA_2$ and MMMR. As revealed by Figure 1, there is a positive relationship between $RMSEA_2$ and MMMR. Figure 1 also demonstrated that most of the MMMR were

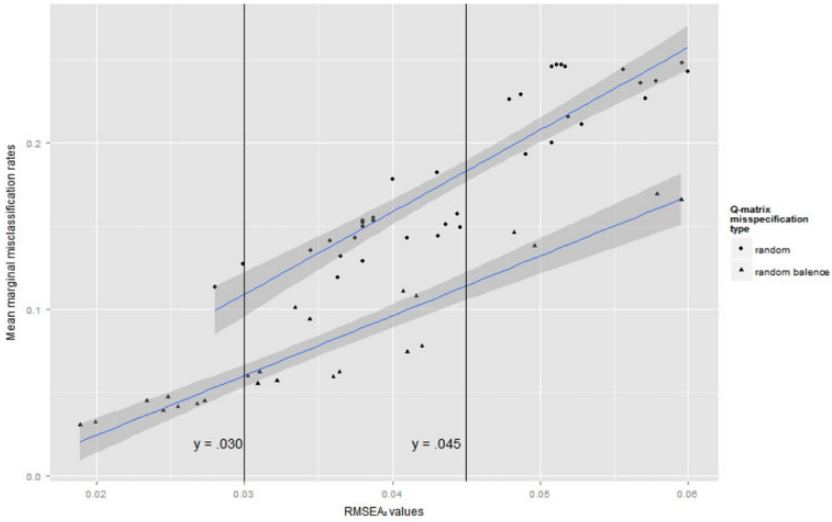

FIGURE 1. *Scatter plot of root mean square error of approximation (RMSEA$_2$) and mean marginal misclassification rates.*

greater than .160 when RMSEA$_2$ $\geq$.045, and most of the MMMR were less than .05, when RMSEA$_2$ $\leq$.030. Thus, we propose that .045 might be a reasonable cutoff criterion of good fit for LCDM, and .030 might be the proper cutoff for excellent fit. Figure 1 also reveals that the types of Q-matrix misspecification had a noticeable impact on the relationship between RMSEA$_2$ and MMMR.

We also found that the MMMR could be well predicted from RMSEA$_2$ values when a linear regression model was used. The coefficients of determination R^2 for the random design and random balance design were .830 and .880, respectively. The correlation was higher in the former than that in the latter. This is probably because there was at least one attribute being correctly specified for each item in the latter.

In Figure 2, in order to systematically illustrate the relationship between MMMR and RMSEA$_2$ under different conditions, MMMR is displayed as a function of RMSEA$_2$ value, attribute correlation, test length, and Q-matrix misspecification type. The sample sizes did not have a noticeable impact on the wrong classification rates, thus they were not included in Figure 2. In terms of the attribute correlations, the classification inaccuracy and RMSEA$_2$ were higher in conditions with $\rho = .5$ than that with $\rho = .8$. For Q-matrix misspecification types, the classification accuracy was higher under the random balance design than those under random design. When $\rho = .8$, the classification accuracy increased with test length. The same trend can also be observed in conditions with random balance design and $\rho = .5$. It's important to point out that even though the rejection rates were 1, the classification accuracy was above .950, when $J = 50$, and $\rho = .8$, under the random balance design. We also found that

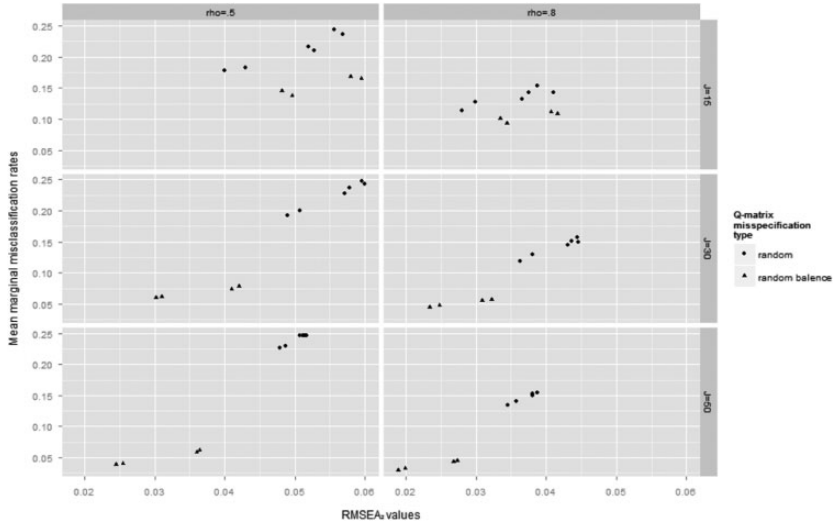

FIGURE 2. *Scatter plot of root mean square error of approximation ($RMSEA_2$) as a function of attribute correlation, test length, and Q-matrix misspecification type.*

the impact of the percentage of Q-matrix misspecification on the MMMR depended on test length, Q-matrix misspecification type, and attribute correlation.

Empirical Illustration

We used fraction subtraction data (Tatsuoka, 1990) to illustrate the usefulness of M_2 and $RMSEA_2$. Although many researchers have studied the data (e.g., DeCarlo, 2010; de la Torre, 2009, 2011; de la Torre & Douglas, 2008; Henson et al., 2009), none could come to an agreement on which model and the corresponding Q-matrix fit the data exactly; in this study, for illustrative purposes, we used M_2 and $RMSEA_2$ statistics to investigate the performance of CDMs among the deterministic inputs, noisy "and" gate model (DINA; de la Torre, 2009; Junker & Sijtsma, 2001), the compensatory reparameterized unified model (C-RUM; Hartz, 2002), and LCDM regarding fraction subtraction data.

A subset of this data set containing the responses of 536 middle school students to 10 fraction subtraction items was analyzed as an example (DeCarlo, 2010; Henson et al., 2009). The items and Q-matrix we implemented were proposed by Henson et al. (2009) and are presented in Table 9. We use the item numbering originally used by Henson et al. (2009). Item 8 was removed from the original data set, as Henson et al. (2009) mistakenly took the test Item 8 $4\frac{5}{7} - 1\frac{4}{7}$ as $4\frac{5}{7} - 1\frac{7}{4}$ (Tatsuoka, 1990). Item 11 was also removed, since different strategies might be used to solve it (Henson et al., 2009).

TABLE 9.
The Q-Matrix for Fraction Subtraction Data.

Item		α_1	α_2	α_3
1	$3\frac{1}{2} - 2\frac{3}{2}$	1	1	0
2	$3 - 2\frac{1}{5}$	1	0	1
3	$3\frac{7}{8} - 2$	1	0	1
4	$4\frac{4}{12} - 2\frac{7}{12}$	1	0	0
5	$4\frac{1}{3} - 2\frac{4}{3}$	1	1	0
6	$\frac{11}{8} - \frac{1}{8}$	1	1	0
7	$2 - \frac{1}{3}$	1	0	1
9	$7\frac{3}{5} - \frac{4}{5}$	1	0	0
10	$4\frac{1}{10} - 2\frac{8}{10}$	1	0	0
12	$4\frac{1}{3} - 1\frac{5}{3}$	1	1	0

Note. We use the items numbers from Henson, Templin, and Willse (2009).

In order to resolve the identification problems and ensure the parameters were reasonable, the item parameters were all estimated by using MMLE algorithm, using flexMIRT (Cai, 2013), and a β (1.6, 1) prior distribution on the item parameters of LCDM and C-RUM was adopted (Bock, Gibbons, & Muraki, 1988). The item parameter estimates of LCDM, DINA, and C-RUM are shown in Table 10. Table 11 provides the values of M_2 and RMSEA_2 for these models.

According to Table 11, the p value of LCDM was .786, which suggests that the saturated LCDM provided a good fit to fraction subtraction data. However, the p values of DINA and C-RUM were significantly smaller than .001, which suggests that the two models did not fit the data well. Following the rules of thumb for RMSEA_2 presented in the section on the performance of RMSEA_2, the DINA and C-RUM did not have a close fit to the data. Among the three fitted models, LCDM was the best choice.

Discussion

Since model–data misfit may yield misleading information, assessing the overall goodness-of-fit for CDMs is of great importance for practitioners. The existing methods for evaluating overall model–data fit under a sparse contingency table have their pros and cons. By contrast, the M_2 statistic developed here hold an obvious predominance in computation efficiency and accurate empirical Type I error rates. In this article, we have conducted simulation studies to compare the performance of χ^2, G^2, and M_2, under the null condition of correct model specifications and the empirical evidence suggested that χ^2, G^2, and M_2 can be

TABLE 10.
Item Parameter Estimates for Fraction Subtraction Data

	LCDM				DINA		C-RUM		
Item	$\lambda_{j,0}$	$\lambda_{j,1,(k)}$	$\lambda_{j,1,(k')}$	$\lambda_{j,2,(k,k')}$	$\lambda_{j,0}$	$\lambda_{j,2,(k,k')}$	$\lambda_{j,0}$	$\lambda_{j,1,(k)}$	$\lambda_{j,1,(k')}$
1	−0.949	−0.539	−0.676	5.490	−1.287	3.277	−1.429	3.017	.846
2	−3.276	0.047	3.737	1.405	−1.878	3.644	−3.249	2.077	3.930
3	0.137	−0.774	0.157	2.475	0.063	1.855	−0.001	0.487	1.203
4	−3.219	4.440			−3.247	4.467	−3.250	4.472	
5	−2.952	4.163	0.666	1.045	−2.579	5.154	−2.626	5.030	.335
6	−3.397	4.724	7.104	−4.827	0.120	2.930	−0.110	2.481	1.860
7	−3.940	2.971	5.508	−2.143	−1.641	4.478	−3.237	2.802	4.697
9	−3.149	5.111			−3.161	5.112	−3.176	5.137	
10	−1.861	3.569			−1.862	3.559	−1.868	3.573	
12	−6.553	8.768	2.256	−2.970	−4.848	6.401	−5.039	6.144	1.029

Note. $\lambda_{j,0}$ is the intercept parameter. $\lambda_{j,1,(k)}$ is the kth main effect parameter. $\lambda_{j,1,(k')}$ is the k'th main effect parameter, $k \neq k'$. $\lambda_{j,2,(k,k')}$ is the two-way interaction parameter. LCDM = log-linear cognitive diagnosis model. DINA = deterministic inputs, noisy "and" gate model. C-RUM = compensatory reparameterized unified model.

TABLE 11.
M_2 *and* $RMSEA_2$ *Statistics for Fraction Subtraction Data Set*

Model	M_2	*df*	*p*	$RMSEA_2$	90% CI
LCDM	9.667	14	.786	0	[0, .028]
DINA	68.369	28	.000	.052	[.036, .068]
C-RUM	54.948	21	.000	.055	[.037, .073]

Note. RMSEA = root mean square error of approximation. CI = confidence interval. LCDM = log-linear cognitive diagnosis model. DINA = deterministic inputs, noisy "and" gate model. C-RUM = compensatory reparameterized unified model; *df* = the degrees of freedom. *p* Values < .001 are reported as .000.

used to test the overall goodness of fit when the contingency table was not sparse; however, only M_2 had accurate Type I error rates when the contingency table was sparse. We also have examined the empirical behavior of M_2 under the incorrect specification of the Q-matrix conditions, and simulation results suggested that M_2 was a powerful tool to detect the misspecification of the model. So we recommend the use of M_2 statistic for assessing the overall exact model fit in CDMs.

The overall goodness-of-fit index only provided information about whether the CDMs and the corresponding Q-matrices fit the data or not. It is also important to assess the goodness of approximation of CDMs. After examining the performance of $RMSEA_2$ under different conditions, especially the

relationship between $RMSEA_2$ and the MMMR, we have found significant correlations between them. The results of simulation studies provided evidence that the cutoff values .030 and .045 might be reasonable criteria for excellent and good fit for LCDM. The results of the simulation study also revealed that the variables of test length, percentage of Q-matrix misspecification, Q-matrix misspecification type, and attribute correlation had a noticeable impact on the classification accuracy and $RMSEA_2$ values. More importantly, although the rejection rates were very high, the classification inaccuracy and $RMSEA_2$ were very low in conditions with $J = 50$ and random balance Q-matrix misspecifications. We thus propose the use of $RMSEA_2$ statistic for assessing the approximate fit in CDMs.

As a widely used parameter estimation method in CDMs (de la Torre, 2011; Robitzsch, Kiefer, George, & Uenlue, 2015), MMLE may yield implausible parameter values for CDMs. When this occurs, it is necessary to impose a reasonable prior on the item parameters (Mislevy, 1986). In the empirical illustration, in order to obtain reasonable item parameter values, model parameters were estimated using flexMIRT, which provides a flexible and practical model parameter estimation procedure in CDMs, and a β (1.6, 1) prior was imposed on the item parameters of LCDM and C-RUM. It is important to point out that these item parameter estimates were still asymptotically normally distributed (Mislevy, 1986). By using the parameter estimates, we assessed the overall goodness of fit and goodness of approximation fit of DINA, C-RUM, and LCDM to fraction subtraction data. The value of M_2 suggested that the saturated LCDM fit the data exactly, however, DINA and C-RUM did not have exact fit; and the values of $RMSEA_2$ suggested that DINA and C-RUM did not have close fit to the data.

There are some limitations in the present study. For example, we only investigated the power of M_2 under the Q-matrix misspecification condition. Further studies are needed to investigate the performance of M_2 under different conditions. The source of misfit at the item level should be investigated when the model does not fit the data well (Liu & Maydeu-Olivares, 2014). Additionally, although the performance of $RMSEA_2$ in LCDM have been investigated in this study, it is desirable to study the performance of $RMSEA_2$ in different CDMs.

Appendix

TABLE A1.
The Q-Matrix for J = *15, 30, and 50,* K = *5*

Item	α_1	α_2	α_3	α_4	α_5	Item	α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	26	1	1	0	1	0
2	0	1	0	0	0	27	1	0	1	1	0
3	0	0	1	0	0	28	1	0	1	0	1
4	0	0	0	1	0	29	0	1	1	0	1
5	0	0	0	0	1	30	0	1	0	1	1
6	1	1	0	0	0	31	1	1	0	0	0
7	1	0	0	0	1	32	1	0	0	0	1
8	0	1	1	0	0	33	0	1	1	0	0
9	0	0	1	1	0	34	0	0	1	1	0
10	0	0	0	1	1	35	0	0	0	1	1
11	1	1	1	0	0	36	1	1	1	0	0
12	1	1	0	0	1	37	1	1	0	0	1
13	1	0	0	1	1	38	1	0	0	1	1
14	0	1	1	1	0	39	0	1	1	1	0
15	0	0	1	1	1	40	0	0	1	1	1
16	1	0	0	0	0	41	1	0	1	0	0
17	0	1	0	0	0	42	1	0	0	1	0
18	0	0	1	0	0	43	0	1	0	1	0
19	0	0	0	1	0	44	0	1	0	0	1
20	0	0	0	0	1	45	0	0	1	0	1
21	1	0	1	0	0	46	1	0	1	1	0
22	1	0	0	1	0	47	1	0	1	0	1
23	0	1	0	1	0	48	1	0	0	1	1
24	0	1	0	0	1	49	0	1	1	0	1
25	0	0	1	0	1	50	0	1	0	1	1

Note. The 15-item test used Items 1–15 and the 30-item test used Items 1–30.

Acknowledgments

The authors would like to thank the editor, Dr. Dan McCaffrey, and several anonymous reviewers for their valuable comments and suggestions, which led to many improvements.

Authors' Note

The views expressed here belong to the authors and do not reflect the views or policies of the funding agencies.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Natural Science Foundation of China (Grant No. 31371047) and by the Fundamental Research Funds for the Central Universities (Grant No. SKZZX2013028).

References

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, *12*, 261–280.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

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. (2013). *flexMIRT® version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]*. Chapel Hill, NC: Vector Psychometric Group.

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.

Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, *50*, 123–140.

DeCarlo, L. T. (2010). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

de la Torre, J., & Douglas, J. A. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

de la Torre, J., & Lee, Y.-S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement*, *50*, 355–373.

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

Henson, R. A., 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.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453.

Junker, B. W., & 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. (2014). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics* (Unpublished doctoral dissertation). Department of Graduate Psychology, James Madison University, Harrisonburg.

Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics*, *18*, 75–86.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*, 336–344.

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.

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, *49*, 354–371.

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. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305–328.

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

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

R Core Team. (2014). *R: A language and environment for statistical computing [Computer software]*. Vienna, Austria: *R* Foundation for Statistical Computing.

Reiser, M. (1996). Analysis of residuals for the multionmial item response model. *Psychometrika*, *61*, 509–528.

Reiser, M., & Lin, Y. (1999). A goodness-of-fit test for the latent class model when expected frequencies are small. *Sociological Methodology*, *29*, 81–111.

Robitzsch, A., Kiefer, T., George, A. C., & Uenlue, A. (2015). *CDM: Cognitive diagnostic modeling [Computer software]*. Retrieved from http://CRAN.R-project.org/package=CDM (R package version 4.2-12)

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., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford.

Sinharay, S. (2006). Model diagnostics for Bayesian networks. *Journal of Educational and Behavioral Statistics*, *31*, 1–33.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models a case study. *Educational and Psychological Measurement*, *67*, 239–257.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*, 29–40.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring Skills and Knowledge Acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.

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

Authors

YANLOU LIU is a PhD student at the School of Psychology at Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing, China, 100875; email: liuyanlou@163.com. His research interests are cognitive diagnostic modeling, item response theory, and structural equation modeling.

WEI TIAN is an assistant professor at Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing, China, 100875; email: tianwei65396@163.com. His research interests include cognitive diagnostic modeling, item response theory, and statistical computing.

TAO XIN is a professor at the Institute of Developmental Psychology and Collaborative Innovation Center of Assessment toward Basic Education Quality at Beijing Normal University, No. 19, XinJieKouWai St., HaiDian District, Beijing, China, 100875; email: xintao@bnu.edu.cn. His research interests are cognitive diagnostic modeling, item response theory, and latent trait models. He is the corresponding author of this article.

Manuscript received July 15, 2014
Revision received June 30, 2015
Accepted September 7, 2015