Article

Applied Psychological Measurement
2017, Vol. 41(8) 632–644

DOI: 10.1177/0146621617710464
journals.sagepub.com/home/apm
SAGE

Investigating the Behaviors of M_2 and $RMSEA_2$ in Fitting a Unidimensional Model to Multidimensional Data

Jie Xu[1], Insu Paek[1], and Yan Xia[2]

Abstract

It has been widely known that the Type I error rates of goodness-of-fit tests using full information test statistics, such as Pearson's test statistic χ^2 and the likelihood ratio test statistic G^2, are problematic when data are sparse. Under such conditions, the limited information goodness-of-fit test statistic M_2 is recommended in model fit assessment for models with binary response data. A simulation study was conducted to investigate the power and Type I error rate of M_2 in fitting unidimensional models to many different types of multidimensional data. As an additional interest, the behavior of $RMSEA_2$ was also examined, which is the root mean square error approximation (RMSEA) based on M_2. Findings from the current study showed that M_2 and $RMSEA_2$ are sensitive in detecting the misfits due to varying slope parameters, the bifactor structure, and the partially (or completely) simple structure for multidimensional data, but not the misfits due to the within-item multidimensional structures.

Keywords

item response theory, limited information statistic, multidimensional structures, M_2

Obtaining the benefits of item response theory (IRT) assumes adequate model-data fit. In this sense, model fit assessment is a fundamental and critical issue in applications of IRT models. Model fit can be examined at two levels in the context of IRT: the item level and the test level. Item level fit analysis assesses model-data fit at the individual item level, and some popular item fit indices include S-χ^2 and S-G^2 statistics (Orlando & Thissen, 2000, 2003). The examination of item fit can diagnose the misfit of items and such item misfit information can be used to improve the overall model fit. Zhang and Stone (2008) describe that the item fit investigation is ''an important complement to model-data fit at the test level'' (p. 182). Test-level fit analysis directly examines the overall model-data fit for the whole test. Several overall model fit or omnibus goodness-of-fit (GOF) statistics have been proposed in the literature. These GOF statistics can be categorized into two main types: full information test statistics (Koehler & Larntz,

[1]Florida State University, Tallahassee, FL, USA
[2]Arizona State University, Tempe, AZ, USA

Corresponding Author:
Jie Xu, Measurement & Statistics Program, Department of Educational Psychology & Learning Systems, Florida State University, 3207 Stone Building, 1114 W. Call St., Tallahassee, FL 32306-4453, USA.
Email: jx12@my.fsu.edu

1980) and limited information test statistics (e.g., Cai & Hansen, 2013; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares, 2013; Maydeu-Olivares & Joe, 2005).

Pearson's test statistic χ^2 and the likelihood ratio test statistic G^2 are the two most well-known full information GOF statistics in IRT applications. When the model parameters are estimated using maximum likelihood (ML), χ^2 and G^2 statistics are asymptotically equivalent and follow a chi-square distribution under the null hypothesis (Koehler & Larntz, 1980; Maydeu-Olivares & Joe, 2005, 2006). However, when the number of items is large, which results in many empty cells in the frequency table having a total of 2^n cells for dichotomous item response data from the n-item test, the Type I error rates of χ^2 or G^2 tend to be incorrect (Cai et al., 2006; Maydeu-Olivares & García-Forero, 2010; Maydeu-Olivares & Joe, 2005, 2006). Therefore, due to data sparseness, these statistics can be applicable only for a test with a small number of items.

Given the limitation of full information testing when data are sparse, three strategies have been proposed to obtain more accurate empirical Type I error rates: (a) pooling cells (Bartholomew & Tzamourani, 1999); (b) resampling methods, such as a parametric bootstrap (Bartholomew & Tzamourani, 1999; Collins, Fidler, Wugalter, & Long, 1993; Tollenaar & Mooijaart, 2003); and (c) limited information testing (e.g., Bartholomew & Leung, 2002; Cai et al., 2006; Maydeu-Olivares & Joe, 2005). Even though pooling cells can reduce the degree of sparseness, this method may result in a loss of information about model misfit (Cai et al., 2006). Time-consuming is the most obvious drawback of the resampling method. Compared with these two methods, limited information testing is much more efficient and thus has been receiving increasing attention. Limited information statistics do not use all information in the data, but low-order margins only. As such they are called ''limited information'' statistics.

Maydeu-Olivares and Joe (2005) proposed a family of limited information statistics, M_r, for models with binary response data. M_r follows asymptotically a central chi-square distribution under the null hypothesis with asymptotically normal consistent estimators. Its degrees of freedom is equal to the number of used multivariate moments (or the number of the margins up to r) minus the number of model parameters. When $r = n$, M_r is the usual chi-square GOF statistic. Maydeu-Olivares and Joe (2005) examined the distributional characteristics of M_2, M_3, and Pearson's χ^2 fitting a two-parameter logistic (2PL) model, with sample sizes of 100 and 1,000, and the number of items equal to 5 and 8. Their results showed that as sparseness increased, the empirical Type I error rates of M_2 remained accurate, while those of χ^2 and M_3 departed from their expected rates. By comparing the 2PL model versus the one-parameter logistic (1PL) model, they also found that M_r for small r (i.e., M_2 and M_3) were more powerful than χ^2. M_2 showed the most accurate empirical Type I error rate and the greatest power among the three test statistics.

Even though limited information testing has received increasing attention in recent years, most studies focus on the power of detecting model misfits of the unidimensional IRT models, especially the 1PL (or Rasch) and 2PL IRT models. However, as far as the authors are aware, no systematic evaluation of M_2 exists for the unidimensional three-parameter logistic (3PL) model. Because of the pseudoguessing parameter in the 3PL model, it has been popular in educational achievement testing which often uses multiple-choice items. In addition to guessing by low-ability examinees, aberrant responses may also appear due to anxiety or careless errors during test. The four-parameter logistic (4PL) model (Barton & Lord, 1981) posits the upper asymptote of the item response function (IRF) to reduce the estimation error for high-ability students who make incorrect responses due to stress, anxiety, carelessness, or distraction. Liao, Ho, Yen, and Cheng (2012) simulated four computerized adaptive tests (CATs) to compare the performance of the 3PL and 4PL IRT models, and they found that ''the 4PL IRT model provides a more robust technique for ability estimation than the 3PL model'' (Liao et al., 2012, p. 1693). In the present study, for unidimensional cases, these two models were considered to evaluate the performance of M_2 in detecting model misspecifications.

Although unidimensionality is a popular assumption in most IRT model applications (Hambleton & Murray, 1983), it may be hard to hold in practice. For example, many standardized tests are constructed based on subcomponents measuring different traits (Ansley & Forsyth, 1985) and ''modern practice often requires examination of single performances from multiple perspectives'' (Adams, Wilson, & Wang, 1997, p. 1). However, research on examining the performance of limited information statistics in detecting misfit due to the multidimensionality of data is still very rare. Cai and Hansen (2013) investigated the performance of M_2 and their newly proposed M_2* for polytomous item response data in testing of hierarchical item factor model (bifactor model with one general dimension and three group-specific dimensions in their study). Even though Cai and Hansen's study reached out to multidimensional data in their evaluation of the limited information tests, their study was limited to only the bifactor structure in their multidimensionality simulation. Thus, the performance of M_2 in many different types of multidimensional structures is still of interest.

M_2 is an overall model fit test. As a statistical test, it becomes very sensitive or too powerful in terms of rejecting the fitted model as sample size increases, although the degree of misfit may be practically tolerable or trivial. The use of an effect size complements the statistical inferential approach for evaluating the model-data fit. Root mean square error approximation (RMSEA; Steiger & Lind, 1980) shows a degree of the model-data misfit and can be considered as an effect size measure. RMSEA was first introduced in structural equation modeling literature and recently it has been applied to the GOF assessment in IRT. Maydeu-Olivares and Joe (2014) addressed RMSEA_2 and RMSEA_n in their study. RMSEA_2 is the RMSEA based on M_2, whereas RMSEA_n refers to the RMSEA based on Pearson's χ^2. In this study, the terminology RMSEA_2 was adopted for the RMSEA calculated based on M_2. Maydeu-Olivares and Joe (2014) investigated the population values of RMSEA_2 and RMSEA_n under different model misspecification and model size conditions, and proposed cutoff criteria of ''adequate fit'' (.089), ''close fit'' (.05), and ''excellent fit'' (.05 / $[k - 1]$, where k is the number of categories) using RMSEA_2 based on their simulation. Although Maydeu-Olivares and Joe's study focused on RMSEA_2, their simulation was restricted to a bidimensional 3PL model, leaving considerable room for research on assessing RMSEA_2 under various multidimensional structures.

Considering alternative multidimensional test structures encountered in practice and also to address the limitations mentioned in the previous studies on M_2, this study aimed to investigate the behaviors of M_2 with regard to its Type I error rates and power for detecting model misfit for multidimensionality in dichotomously scored item response data when a unidimensional model was fit, such as the 1PL, 2PL, and 3PL models. The data-generating models in the current study included both of the unidimensional IRT models (i.e., the 2PL, 3PL, and 4PL models) and diverse multidimensional IRT (MIRT) models (i.e., simple structure, partially simple structure, completely cross-loading structure, bifactor, and noncompensatory MIRT models). Because the calculation of RMSEA_2 is very straightforward once M_2 is obtained, as an additional interest, the behavior of RMSEA_2 was also examined under the same simulated conditions. Note again that the performances of M_2 and RMSEA_2 in these diverse alternative data structures have not been evaluated in the current literature. The results from the present study, therefore, could benefit practitioners and researchers who would like to use M_2 and RMSEA_2 in assessing model-data misfit in IRT applications.

Method

M_2 and RMSEA_2

The M_2 statistic has the following form:

$$M_2 = N\hat{\mathbf{e}}_2^{'}\hat{\mathbf{C}}_2\hat{\mathbf{e}}_2, \quad (1)$$

where N is the sample size, $\hat{\mathbf{e}}_2$ is the vector consisting of the estimated first- and second-order residual proportions, and $\hat{\mathbf{C}}_2$ is a weight matrix that involves an asymptotic covariance matrix of the first- and second-order residual proportions and a matrix of derivatives of the model-implied marginal probabilities up to the second order with respect to the model parameters. See Maydeu-Olivares and Joe (2005, 2006) for more details.

The value of RMSEA_2 is calculated by the following equation:

$$\text{RMSEA}_2 = \sqrt{\text{Max}\left(\frac{M_2 - df}{N \times df}, 0\right)}, \quad (2)$$

where M_2 is the M_2 in Equation 1 and df is the degrees of freedom for M_2. Refer to Maydeu-Olivares, Cai, and Hernández (2011) and Maydeu-Olivares and Joe (2014) for the details of the RMSEA_2 and its confidence interval calculation.

Study Design

Different types of unidimensional and MIRT models were considered for data generation. The unidimensional models were 2PL, 3PL, and 4PL models mentioned earlier. MIRT models were either compensatory or noncompensatory. For compensatory MIRT models, a low ability in one dimension can be fully compensated by a high ability in another dimension to reach a high probability of a correct answer. In Noncompensatory MIRT models, a high ability in one dimension can only partially compensate a low ability in another dimension (i.e., a deficiency in one dimension cannot be offset much by an increase in others). Compensatory models are more popular in the IRT literature. However, noncompensatory MIRT models might be more appropriate for a test having items with conjunctive component processes (Maris, 1999). Both noncompensatory and compensatory MIRT models were included in this simulation. See Bolt and Lall (2003) and Reckase, 2009 for details of compensatory and noncompensatory MIRT models.

Four types of compensatory MIRT models were considered for data generation: (a) completely cross-loading structure (or equivalently within-item multidimensionality), indicating that each of the items on the test measures more than one dimension; (b) multidimensional between-item structure (or equivalently simple structure), indicating that a test contains multiple unidimensional subscales; (c) partially cross-loading structure (or equivalently partially simple structure), meaning that there exists some items measuring a single dimension while other items measure multiple dimensions; and (d) bifactor structure which contains a general dimension and several specific dimensions.

To be specific, for the completely cross-loading structure, a two-dimensional within-item (2DW) model (e.g., a physics test having relatively long wordy item stems for which reading ability may play an important role) was adopted for data generation in the current study. For the simple structure, a four-dimensional simple structure (4DS) model was used (e.g., a math test containing algebra, probability, statistics, and geometry). As for the partially simple structure, a two-dimensional partially simple structure model (2DPS) was used, where one set of items measure the first dimension, another set of items measure both the first and second dimensions, and the third set of item measure the second dimension. The bifactor models included in the study were five-dimensional with one general dimension and four specific dimensions (5DBi).

For the noncompensatory models, two-dimensional noncompensatory (2DNC) MIRT models were used.

Table 1. Description of the Naming Convention for Multidimensional Data Structures.

Multidimensional data structures	
Compensatory	
2DW_2PL	Two-dimensional within-item 2PL model
2DW_3PL	Two-dimensional within-item 3PL model
2DPS_2PL	Two-dimensional partially simple structure 2PL model
2DPS_3PL	Two-dimensional partially simple structure 3PL model
4DS_2PL	Four-dimensional simple structure 2PL model
4DS_3PL	Four-dimensional simple structure 3PL mode
5DBi_2PL	Five-dimensional bifactor 2PL model
5DBi_3PL	Five-dimensional bifactor 3PL model
Noncompensatory	
2DNC_2PL	Two-dimensional noncompensatory 2PL model
2DNC_3PL	Two-dimensional noncompensatory 3PL model

Note. 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model.

The convention used for naming multidimensional data structures in this study followed the form of ''xDyy_zPL'' (x: the number of dimensions, yy: the type of multidimensionality, z: the number of item parameters in the calibrating model). Table 1 provides the labels of all the multidimensional models used for data generation.

In total, 13 latent trait structures were simulated: three unidimensional structures (i.e., the 2PL, 3PL, and 4PL models) and 10 multidimensional structures (2DNC_2PL/3PL models, 2DW_2PL/3PL model, 2DPS_2PL/3PL models, 4DS_2PL/3PL model, and 5DBi_2PL/3PL models). In addition to different latent trait structures, two levels of sample size (i.e., 750 and 1500) and two levels of test length (i.e., 20 and 40) were used to generate data. A total of 52 conditions were considered in data generation, created by fully crossing the three factors (2 sample sizes $\times$ 2 test lengths $\times$ 13 latent trait structures). For each condition, 500 data sets were replicated, resulting in a total of 26,000 (500 $\times$ 52) data sets. The 1PL, 2PL, and 3PL models were fitted to each data set. M_2 statistic and its corresponding p value, and RMSEA_2 values were obtained from the three models fitted to each data set.

MIRT Models for Data Simulation

The general form of the compensatory IRF used for data simulation was

$$P_j(\theta_i) = g_j + (d_j - g_j)\frac{\exp\left(\sum_k a_{jk}\theta_{ik} - b_j\right)}{1 + \exp\left(\sum_k a_{jk}\theta_{ik} - b_j\right)}, \tag{3}$$

where θ_i is an ability vector for person i; θ_{ik} is person i's latent trait (or ability) for dimension k; a_{jk} is the item discrimination or item slope of item j for dimension k; b_j is the intercept for item j (b_j/a_j is the item difficulty parameter for item j when $k = 1$); g_j is the (pseudo-)guessing parameter for item j representing the IRF's lower bound or asymptotic; and d_j is the upper bound or asymptotic. The 2DW model was obtained when $k = 2$, $d_j = 1$, and $g_j = 0$ (2PL), and when $k = 2$ and $d_j = 1$ (3PL). The 2DPS_2PL (or _3PL) model was obtained when $k = 2$, $d_j = 1$, and $g_j = 0$ (or $k = 2$ and $d_j = 1$) with the constraint $a_{jk} = 0$ for some j and k such that one item set measures the first dimension, another set measures the second dimension, and the third set measures both

dimensions. The 4DS_2PL (or _3PL) model was obtained when $k = 4$, $d_j = 1$, and $g_j = 0$ (or $k = 4$ and $d_j = 1$) with the constraint $a_{jk} = 0$ for some j and k such that each of the four different item sets measured its own single dimension. The 5DBi_2PL (or _3PL) model had $k = 5$, $d_i = 1$, and $g_j = 0$ (or $k = 5$ and $d_j = 1$) with the constraint $a_{jk} = 0$ for some j and k such that each item measured two dimensions where the first dimension was a general dimension and the second dimension is a specific dimension.

For the simple structure (i.e., 4DS_2PL/4DS_3PL) used in this study, the number of items loading on each dimension was identical. For example, for the test length of 40 items, there were 10 items loading on each dimension. For the bifactor structure (i.e., 5DBi_2PL/5DBi _3PL), the number of items loading on each of the specific dimension was also identical. For instance, 10 items loaded onto each of the four specific dimensions in a 40-item test condition. Under the partially simple structure conditions (i.e., 2DPS_2PL/2DPS_3PL), with the test length of 40 items, 13 items (Items 1 through 13) loaded onto the first dimension only; 14 items (Items 14 through 27) loaded onto both dimensions; and the rest of 13 items (Items 28 through 40) loaded onto the second dimension only. In the same partially simple structure with a 20-item condition, six items (Items 1 through 6) loaded onto the first dimension only; eight items (Items 7 through 14) loaded onto both dimensions; and the rest of six items (Items 15 through 20) loaded onto the second dimension only.

A noncompensatory MIRT model ''separates the cognitive tasks in a test item into parts and uses a unidimensional model for each part'' (Reckase, 2009, p. 79), which yields a nonlinear combination of different latent dimensions. Below is the IRF for the 2DNC_3PL model. Again, the IRF for the 2DNC_2PL model can be obtained if $g_j = 0$.

$$P_j(\theta_i) = \prod_{k=1}^{2}\left[g_j + \left(1 - g_j\right)\frac{\exp\left(a_{jk}\theta_{ik} - b_{jk}\right)}{1 + \exp\left(a_{jk}\theta_{ik} - b_{jk}\right)}\right], \tag{4}$$

where b_{jk} is the jth item intercept in the kth dimension.

Model Parameter Values for Data Simulation

For unidimensional models, a scalar of person latent trait parameter was drawn from the standard normal distribution. For multidimensional models, a vector of the person latent trait parameters was drawn from a multivariate normal distribution with the zero mean vector and the covariance matrix of $\boldsymbol{\Sigma}$ that has 1s in the diagonals (the correlation in $\boldsymbol{\Sigma}$ is described later.) The intercept parameters (b) were randomly generated from a standard normal distribution and constrained within the range of -2 to $+2$. Considering that standardized achievement tests are designed to measure a dominant trait, the primary dimension should demonstrate a higher difficulty than the second dimension if there is any (Ansley & Forsyth, 1985; Li & Olejnik, 1997). Based on this conceptualization, intercepts on the second dimension were set by subtracting .5 from that of the primary dimension. However, for the 2DPS models, intercepts for both dimensions were randomly generated from the target distribution, considering that these two dimensions are equally important. The slope parameters were randomly generated from the lognormal distribution with a mean of 0 and a standard deviation of .5, and truncated within [.5, 4]. For MIRT models, slopes on the second dimension were reduced by .3 from the primary dimension. Slopes on each dimension in the 2DPS models were generated separately from the lognormal distribution. The lower asymptotes were randomly generated from a uniform distribution (0, .3); the upper asymptotes were randomly generated from a uniform distribution (.8, 1). For each simulation model in each replication, item parameters were regenerated following the same

procedures. Note that every data set was generated by a different set of item parameter values. This varying-parameter set in each replication instead of a fixed parameter set for a test ensured more realistic simulation and wider scope of generality than the fixed item parameter simulation in the sense that tests can have different item parameter values even if they share the same test length and measure the same construct(s). For MIRT models with only two dimensions (the 2DPS, 2DW, and 2DNC models), a moderate correlation, .5, was imposed between the primary and second dimensions. For the 4DS models, correlation among dimensions was specified as either .7 or .5. The more detailed description of simulation IRF forms and distributions used to generate the model parameters is provided in a separate online appendix.

For model estimation, the flexMIRT program (Version 2.0; Cai, 2013) was used with the marginal maximum likelihood estimation (Bock-Aiktin Expectation-Maximization [BAEM] option in flexMIRT) and with the default convergence criteria and the default error covariance matrix calculation method (i.e., the cross-product approximation). To minimize the chance of convergence problems in the 2PL and 3PL model estimation, item parameter priors—lognormal for slope: $a \sim \text{lognormal}(0, .5^2)$ and beta prior for the lower asymptote: $g \sim \text{Beta}(4, 18)$—were used. No prior distribution was used for the intercept parameters.

Results

Convergence for the model estimation was investigated first. The data sets with convergence problems were removed in the analysis. For those data sets without convergence problems, rejection rates of M_2 and the average values of RMSEA_2 were the outcome variables to be analyzed.

Convergence Rates

Convergence rates for 142 of the total 156 conditions (91%) were above the 90%. Those 14 conditions showing lower than 90% convergence were from the 2DW and 2DPS structure models. For two of the 14 conditions, 2PL models were fitted (with the convergence rates from .74 to .80), and for the rest of 12 of the 14 conditions, data were analyzed using the 3PL models (with the convergence rates from .32 to .89). Note that the low convergence by the 3PL fit may be partly due to the difference between the prior distribution used for the guessing parameters in fitting the 3PL model and the distribution used to generate the guessing parameters for data simulation—$g \sim \text{Beta}(4, 18)$ for the 3PL model fitting, and $g \sim \text{uniform}(0, .3)$ for data generation. Detailed convergence rates or valid number of replications for all conditions can be obtained upon request from the authors.)

Type I Error Rate and Power of M_2

In spite of the low convergence rates in those 14 conditions, they showed very similar standard errors of the rejection rates of M_2 as the other conditions. The standard error was calculated by $\sqrt{p(1-p)/R}$, where p is the rejection rate and R is the number of valid/converged runs. The standard errors of the 14 low convergence conditions were about .01 or less. Table 2 displays the rejection rates for the M_2 statistic. The values within rectangular boxes represent the empirical Type I error rates while all the other values are empirical power.

When the hypothesis testing for $H_0 : \pi = .05$, where π is equivalent to the nominal significance level, was conducted for the empirical Type I error rates, the values of .03 and .028 in the 2PL model case and .025 in the 3PL model case resulted in the statistical rejection of H_0 under the .05 significance level (those are italicized in Table 2). Overall, a bit conservative, but

Table 2. Rejection Rates in Terms of the M_2 Statistic.

n	Item	Calibrating model	Data generating model												
			2PL	3PL	4PL	2DNC_2PL	2DNC_3PL	2DW_2PL	2DW_3PL	2DPS_2PL	2DPS_3PL	4DS_2PL	4DS_3PL	5DBi_2PL	5DBi_3PL
750	20	1PL		.998	.978										
		2PL	.032	**.05**	**.062**	**.052**	**.034**	**.027**	**.085**	.996	.96		.982	.998	.98
		3PL	NA	.035	**.046**	**.031**	**.029**	**.013**	**.02**	.995	.933	.998	.944	.998	.965
	40	1PL													
		2PL	.032	**.108**	**.056**	**.07**	**.06**	**.005**	**.101**						
		3PL	NA	.025	**.034**	**.05**	**.041**	**0**	**.009**						
1,500	20	1PL													
		2PL	*.03*	**.078**	**.052**	**.062**	**.044**	**.037**	**.153**						.998
		3PL	NA	.039	**.036**	**.051**	**.042**	**.025**	**.029**						.998
	40	1PL													
		2PL	*.028*	**.124**	**.068**	**.102**	**.066**	**.013**	**.298**						
		3PL	NA	.032	**.032**	**.131**	**.048**	**.011**	**.016**						

Note. NA indicates that values are not available in the table. 2DNC_2PL and 2DNC_3PL are two-dimensional noncompensatory 2PL model and two-dimensional noncompensatory 3PL model, respectively; 2DW_2PL and 2DW_3PL indicate two-dimensional within-item 2PL model and two-dimensional within-item 3PL model, respectively; 2DPS_2PL and 2DPS_3PL denote two-dimensional partially simple structure 2PL model and two-dimensional partially simple structure 3PL model, respectively;4DS_2PL and 4DS_3PL are four-dimensional simple structure 2PL model and four-dimensional simple structure 3PL model, respectively; 5DBi_2PL and 5DBi_3PL indicate five-dimensional bifactor 2PL model and five-dimensional bifactor 3PL model, respectively. 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model.

no Type I error rate was inflated to be greater than .05. In general, these results support what Maydeu-Olivares and Joe (2005) have reported in their study.

The power reached or was nearly 1 when calibrated with the 1PL model no matter what the generated data structures were. However, when the 2PL/3PL unidimensional model was fitted to the data that did not conform to the fitted model, except the three multidimensional cases (2DPS, 4DS, and 5DBi), the power of M_2 decreased by a large degree and it was below .3 (see the bold numbers in Table 2), even though the power increased as the sample size increased. (To facilitate readers' understanding of the power shown by M_2 in Table 2, a separate online appendix provides Figures 1 and 2 for unidimensional and multidimensional data conditions, respectively.) It was clear that M_2 was sensitive to misfits due to varying slope parameters, bifactor structure, and at least partially simple structure for multidimensional data. However, M_2 did not show enough power to detect misfits due to the lower/upper asymptote parameters and the within-item multidimensionality. Overall, the average power of M_2 across all multidimensional conditions (M = .746, SD = .42) was higher than that of unidimensional conditions (i.e., misspecification of IRF form rather than dimensionality; M = .457 and SD = .478). The average power of M_2, when excluding the 2DNC and 2DW models which showed weak power, was nearly 1 (M = .996, SD = 0.0001).

Behaviors of $RMSEA_2$

Values of RMSEA ''provide an assessment of the severity of the model misspecification'' (Cai & Hansen, 2013, p. 271). Table 3 displays the average $RMSEA_2$ values for each simulated condition.

For the description of the results, let ''consistent condition'' be the condition where data are consistent with the fitted model and ''inconsistent condition'' be the condition where data do not follow the fitted model. Also, term ''average'' is omitted in the following description for simplicity. The minimum and maximum $RMSEA_2$ values were .001 and .064, respectively, across all conditions. $RMSEA_2$ values for the fitted 2PL and 3PL models were very low, ranging from .002 to .004 in the consistent conditions. Under the inconsistent conditions, $RMSEA_2$ values were .022 and .024, respectively, for the fitted 2PL and 3PL models. These values are much larger than those from the consistent conditions. When the 1PL model was fitted under the inconsistent conditions, $RMSEA_2$ was .046, which is higher than those for the fitted 2PL and 3PL models under the same inconsistent conditions. Regardless of fitted models, $RMSEA_2$ was .031 under all inconsistent conditions. The conventional rule of thumb known in the structural equation modeling applications for RMSEA (e.g., RMSEA $<$ .05; Browne & Cudeck, 1993; and RMSEA $<$ .06; Hu & Bentler, 1999) is very likely to result in the failure of misfit detection when it is applied to the $RMSEA_2$ results from this study. Figures showing patterns of $RMSEA_2$ values are available in the online appendix.

Maydeu-Olivares and Joe (2014) proposed a cutoff of .05 for close fit for binary response IRT models. Overall, the results in Table 3 appear to suggest a more conservative value than .05 to detect model misfits. Marsh, Hau, and Wen (2004) discussed the issues related to finding rules of thumb for GOF indices, addressing limitations of the use of those rules of thumb or cutoffs. In the results, under the inconsistent conditions, the 1PL model applications resulted in the $RMSEA_2$ value which cannot be treated as similar or the same as that from the 2PL or 3PL model applications. This seems to be consistent with Marsh et al.'s position that a cutoff or rule of thumb in the use of GOF indices should not be regarded as a universal golden rule. Despite this limitation, one might still want to have an approximate threshold for a preliminary check regarding a model misfit. For example, if .03 was applied to the simulated data, 96.6% correct decision can be found for the inconsistent conditions with the 1PL model fitted, 47.2% correct

Table 3. $RMSEA_2$ Values.

			Data generating model												
n	Item	Calibrating model	2PL	3PL	4PL	2DNC_2PL	2DNC_3PL	2DW_2PL	2DW_3PL	2DPS_2PL	2DPS_3PL	4DS_2PL	4DS_3PL	5DBi_2PL	5DBi_3PL
750	20	1PL	.044	.037	.031	.043	.039	.061	.051	.072	.056	.05	.039	.063	.049
		2PL	.004	.005	.005	.005	.005	.003	.006	.042	.029	.042	.03	.051	.037
		3PL	NA	.004	.004	.005	.004	.002	.004	.042	.028	.04	.028	.051	.036
	40	1PL	.035	.031	.026	.035	.032	.048	.041	.063	.048	.049	.037	.062	.047
		2PL	.003	.005	.004	.004	.003	.001	.005	.044	.031	.044	.031	.053	.038
		3PL	NA	.003	.003	.003	.002	.001	.001	.044	.03	.042	.03	.053	.037
1,500	20	1PL	.045	.038	.032	.043	.038	.061	.051	.072	.056	.051	.038	.064	.05
		2PL	.003	.004	.004	.004	.004	.003	.006	.043	.03	.042	.032	.051	.037
		3PL	NA	.003	.003	.004	.004	.003	.003	.043	.029	.04	.028	.051	.037
	40	1PL	.035	.031	.027	.035	.033	.049	.041	.064	.049	.049	.036	.062	.046
		2PL	.002	.004	.003	.003	.002	.001	.006	.045	.031	.044	.031	.053	.038
		3PL	NA	.002	.002	.004	.002	.001	.001	.045	.03	.043	.029	.053	.038

Note. NA indicates that values are not available in the table. 2DNC_2PL and 2DNC_3PL are two-dimensional noncompensatory 2PL model and two-dimensional noncompensatory 3PL model, respectively; 2DW_2PL and 2DW_3PL indicate two-dimensional within-item 2PL model and two-dimensional within-item 3PL model, respectively; 2DPS_2PL and 2DPS_3PL denote two-dimensional partially simple structure 2PL model and two-dimensional partially simple structure 3PL model, respectively;4DS_2PL and 4DS_3PL are four-dimensional simple structure 2PL model and four-dimensional simple structure 3PL model, respectively; 5DBi_2PL and 5DBi_3PL indicate five-dimensional bifactor 2PL model and five-dimensional bifactor 3PL model, respectively. 2PL model = two-parameter logistic model; 3PL model = three-parameter logistic model.

decision in the inconsistent conditions with the 2PL model fitted, and 54.1% correct decision in the inconsistent conditions with the 3PL model fitted. However, when data followed partially (or completely) simple structure (2DPS and 4DS) or bifactor structure (5DBi), the .03 threshold provided strong evidence to detect model misfits when the 2PL or 3PL model was fitted. The percentage of correct decision in these multidimensional data conditions was 93.3% with the 2PL model fitted and 90.4%, with the 3PL model fitted. When data were consistent with the fitted model of the 2PL or 3PL model, the .03 threshold resulted in zero percent of incorrect decision (i.e., not rejecting the fitted model). Note that .03 here was not based on quantitative analysis to obtain an optimal value, but on subjective observation of the results. Also, the model misfit detection rates above are subject to the problem of capitalizing on chance because the .03 value was applied to the same data where it was selected.

Discussion and Conclusion

The main focus of the current study was on evaluating the performance of M_2 in detecting model misspecifications, especially for the IRF form violation due to lower/upper asymptote and for the violation of unidimensionality when multidimensional data were fitted by a unidimensional IRT model. In addition, the behaviors of $RMSEA_2$ for detecting model misspecification were investigated.

M_2 produced accurate Type I error rates for most conditions and all Type I error rates were under the nominal significance rate. It was found that M_2 has large power to detect multidimensionality that is partially (or completely) simple structure or bifactor structure. M_2 was also very powerful to reject the null hypothesis of constant slope in the IRFs when the data were from models having different slopes in the IRFs regardless of unidimensional or multidimensional structure. However, it was not sensitive enough in detecting misfits due to ignoring the nonzero lower asymptote and the upper asymptote less than one, the within-item multidimensional structures, and the noncompensatory multidimensional structures. Part of the study's findings are consistent with results from Cai and Hansen's (2013) study and Hansen, Cai, Monroe, and Li (2014), both of which reported that M_2 was powerful to detect the presence of secondary dimensions in data having a bifactor structure.

The original RMSEA was first proposed by Steiger and Lind (1980) in the context of factor analysis (for continuous variables) to evaluate model–data fit, taking model complexity into account. Maydeu-Olivares and Joe (2014) proposed the use of $RMSEA_2$ for IRT and remarked .05 as the cutoff for a ''close'' fit and .089 as ''adequate'' for binary IRT model–data fit when using $RMSEA_2$. These suggested values do not completely match with the simulation results which indicated a more conservative value than their close or adequate criterion for detecting the violation of unidimensionality (i.e., detecting multidimensionality).

As described previously, when using M_2 and $RMSEA_2$, misfits due to the within-item multidimensional structures cannot be easily detected, but these two statistics are sensitive to misfits due to varying slope parameters, bifactor structure, and partially (or completely) simple structure. For this, it is conjectured that the power of M_2 is partly influenced by the number of items measuring a single dimension or the degree of simple structure. Future studies may examine this hypothesis through systematic variations in the extent of simple structure (and possibly with varying magnitudes of correlations among dimensions) in the data.

Although many different unidimensional and multidimensional data structures were accommodated in this study, other limitations exist. Apparently, all the uni- and multidimensional models used do not represent all of the possible multi- and unidimensional data structures (e.g., a mixture IRT model structure). The use of fixed dimension correlations is another limitation. The between-dimension correlations were fixed to either .5 for the two-dimensional structures

or .7/.5 for the four-dimensional structures. Future studies may explore the performance of the M_2 statistic by extending the limited levels of the factors used in the current study design and incorporating different multidimensional data structures.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplementary material is available for this article online.

References

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

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*, 37-48.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p 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 & Research*, *27*, 525-546.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Service.

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

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

Cai, L. (2013). flexMIRT: A numerical engine for flexible multilevel multidimensional item analysis and test scoring (Version 2.0) [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., Coffmanand, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, *59*, 173-194.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research*, *28*, 375-389.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, Canada: Educational Research Institute of British Columbia.

Hansen, M., Cai, L., Monroe, S., & Li, Z. (2014). *Limited-information goodness-of-fit testing of diagnostic classification item response theory models* (CRESST Report No. 840). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55.

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

Li, M. F., & Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, *21*, 215-231.

Liao, W.-W., Ho, R.-G., Yen, Y.-C., & Cheng, H.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, *40*, 1679-1694.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*, 187-212.

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

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling*, *18*, 333-356.

Maydeu-Olivares, A., & García-Forero, C. (2010). Goodness-of-fit testing. *International Encyclopedia of Education*, *7*, 190-196.

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. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713-732.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*, 305-328. doi:10.1080/00273171.2014.911075

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341.

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

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of S - X2: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*, 289-298.

Reckase, M. D. (2009). *Multidimensional item response theory (Statistics for social and behavioral sciences)*. Dordrecht, The Netherlands: Springer.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically based tests for the number of common factors*. Paper presented at the Psychometrika Society meeting, Iowa City, IA.

Tollenaar, N., & Mooijaart, A. (2003). Type I errors and power of the parametric bootstrap goodness-of-fit test: Full and limited information. *British Journal of Mathematical and Statistical Psychology*, *56*, 271-288.

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, *68*, 181-196.