Assessing Item-Level Fit for Higher Order Item Response Theory Models

Xue Zhang
Northeast Normal University
Chun Wang
University of Washington
Jian Tao
Northeast Normal University

The project is supported by the National Natural Science and Social Science Foundations of China (Grants 11571069) and Institute of Education Sciences grant R305D170042 (originally R305D160010).

Citation: Zhang, X., Wang, C., & Tao, J. (2018). Assessing Item-level fit for higher order item response theory models. *Applied Psychological Measurement, 42,* 644-659.

Article

Applied Psychological Measurement
2018, Vol. 42(8) 644–659

DOI: 10.1177/0146621618762740
journals.sagepub.com/home/apm
SAGE

Assessing Item-Level Fit for Higher Order Item Response Theory Models

Xue Zhang[1], Chun Wang[2] and Jian Tao[1]

Abstract

Testing item-level fit is important in scale development to guide item revision/deletion. Many item-level fit indices have been proposed in literature, yet none of them were directly applicable to an important family of models, namely, the higher order item response theory (HO-IRT) models. In this study, chi-square-based fit indices (i.e., Yen's Q_1, McKinley and Mill's G^2, Orlando and Thissen's $S\text{-}X^2$, and $S\text{-}G^2$) were extended to HO-IRT models. Their performances are evaluated via simulation studies in terms of false positive rates and correct detection rates. The manipulated factors include test structure (i.e., test length and number of dimensions), sample size, level of correlations among dimensions, and the proportion of misfitting items. For misfitting items, the sources of misfit, including the misfitting item response functions, and misspecifying factor structures were also manipulated. The results from simulation studies demonstrate that the $S\text{-}G^2$ is promising for higher order items.

Keywords

higher order IRT models, item fit, $S\text{-}X^2$, $S\text{-}G^2$, false positive rate, correct detection rate

Introduction

Item response theory (IRT) models have gained widespread use since their introduction. Originally, it was assumed that the latent trait is unidimensional (Baker & Kim, 2004). More recently, multidimensional IRT models are presented to relax such an assumption (Reckase, 2009). Nonetheless, neither unidimensional IRT models nor multidimensional IRT models capture the hierarchical nature of latent traits, in which the multiple, domain-level latent traits are related to a higher order general trait. de la Torre and Song (2009) proposed a higher order item response theory (HO-IRT) model that captures the overall and domain-specific abilities. By positing a higher order structure, the HO-IRT model has been shown to measure domain-specific abilities and estimate item parameters better than the typical multidimensional IRT models. In recent years, the HO-IRT model has been used in a wide variety of domains. For instance, it is used to evaluate examinees' abilities in computerized adaptive testing (Huang,

[1]Northeast Normal University, Changchun, Jilin, China
[2]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Jian Tao, Key Laboratory for Applied Statistics, School of Mathematics and Statistics, Northeast Normal University, 5268 Renmin Street, Changchun, Jilin, 130024, China.
Email: taoj@nenu.edu.cn

Chen, & Wang, 2012; Lee, 2014; Wang, 2014); to measure testlet-based items (Huang & Wang, 2013); to evaluate hierarchical latent traits (Huang, Wang, Chen, & Su, 2013); to analyze sparse, multigroup data for integrative data analysis (Huo et al., 2015); to assess longitudinal data (Huang, 2015); to measure academic growth within both IRT and structural equation modeling (SEM) frameworks (Wang, Kohli, & Henn, 2016); and to integrate with mixture IRT models to account for subclasses within a population (Huang, 2017).

When a parametric model is fitted to data, item-level fit is usually assessed to guide item revision/deletion. To evaluate item fit, numerous statistical procedures have been introduced in IRT literature (Bock,1972; Bock & Haberman, 2009; Chon, Lee, & Dunbar, 2010; Demars, 2005; Glas & Suárez-Falcón, 2003; Haberman, 2009; Haberman, Sinharay, & Chon, 2013; Kang & Chen, 2008; LaHuis, Clark, & O'Bruen, 2011; Li & Rupp, 2011; Liang & Wells, 2009; McKinley & Mills, 1985; Muraki & Bock, 1997; Orlando & Thissen, 2000, 2003; Ranger & Kuhn, 2012; Roberts, 2008; Sinharay, 2005, 2006; Stone, 2000; Stone & Zhang, 2003; Suárez-Falcón & Glas, 2003; Wang, Shu, Shang, & Xu, 2015; Wells & Bolt, 2008; Yen, 1981; Zhang & Stone, 2008). Among them, chi-square-based item fit indices (i.e., Q_1, G^2, $S\text{-}X^2$, and $S\text{-}G^2$) are the most popular family of statistical indices and they have been used to examine model misspecification under dichotomous or/and polytomous items (Chon et al., 2010; Kang & Chen, 2008; Yen, 1981; Liang & Wells, 2009; McKinley & Mills,1985; Orlando & Thissen, 2000; Wang et al., 2015). $S\text{-}X^2$ and Q_1 were used to test violation of the monotonicity assumption of the item response function (IRF; Orlando & Thissen, 2003), or to test the item misfit due to model misspecification or Q-matrix misspecification in latent class models (Wang et al., 2015). Stone's χ^{2*} and G^{2*} were used to detect item parameter drift (LaHuis et al., 2011; Stone & Zhang, 2003). Also, $S\text{-}X^2$ was used to identify the factor structure of the test, namely, to differentiate simple from complex multidimensional structure (Zhang & Stone, 2008), or to differentiate unidimensional from bifactor or multidimensional structure (Li & Rupp, 2011).

Furthermore, Haberman (2009) used generalized residuals to assess item fit based on *t* statistics. Different from the previous methods (i.e., chi-square-based statistics or generalized residuals) which used different forms of discrepancy measures to quantify the discrepancy between model prediction and observation, the Lagrange multiplier (LM) test (Glas & Suárez-Falcón, 2003) was proposed to identify item misfit due to violation of local independence. Under Bayesian framework, the posterior predictive model checking (PPMC) method (Sinharay, 2005, 2006) was also used to assess item fit. Different forms of discrepancy measures were considered within the PPMC method framework, such as Yen's Q_1 (Sinharay, 2005; Wang et al., 2015), Stone's χ^{2*} (Wang et al., 2015), and Orlando and Thissen's $S\text{-}X^2$ and $S\text{-}G^2$ (Sinharay, 2006).

Moreover, the root integrated squared error (RISE, Douglas & Cohen, 2001) was presented to investigate the fit of parametric IRT models by comparing them with models fitted under nonparametric assumptions. To test item misfit due to model misspecification, RISE outperformed G^2 and $S\text{-}X^2$ in that it controlled Type I error rates and provided adequate power (Liang & Wells, 2015; Liang, Wells, & Hambleton, 2014; Wells & Bolt, 2008).

On the contrary, the limited information fit statistics (Bartholomew & Leung, 2002; Cai, Maydeu-Olivares, Coffman, & Thissen, 2006; Maydeu-Olivares & Joe, 2005; Reiser, 2008) used the marginal tables (i.e., the cross tabulations of item pairs or item triplets), rather than frequencies of single response patterns in chi-square-based item fit statistics to identify misfit. These sets of indices were often used to assess item fit in sparse contingency tables (Cai et al., 2006; Maydeu-Olivares & Joe, 2005), detect local independence (Liu & Maydeu-Olivares, 2013), and identify the source of misfit (Liu & Maydeu-Olivares, 2014).

While all above-cited literature focused on assessing item-level fit for nonhierarchical models, the main focus of this article is to assess item-level fit for one kind of hierarchical models, that is, HO-IRT models. Li and Rupp (2011) is the only prior study that evaluated the

performance of multivariate $S\text{-}X^2$ with the bifactor model, the multidimensional item response theory (MIRT) model, and the unidimensional item response theory (UIRT) model. However, their study is limited in the following aspects: (a) Only two dimensions were considered, and the performance of $S\text{-}X^2$ with more than two dimensions was not examined; (b) the performance of $S\text{-}X^2$ for comparing the bifactor model with the HO-IRT model was not mentioned; and (c) the recursive algorithm which was needed to compute $S\text{-}X^2$ for hierarchical models was not developed in full generality (Cai, 2015).

The primary goals of this study, thus, are to (a) investigate the performances of four chi-square-based indices: Q_1, G^2, $S\text{-}X^2$, and $S\text{-}G^2$ to detect the item misfit due to misspecified IRFs where the items conform to the HO-IRT model and (b) to examine the performances of the indices to detect item misfit when misfit is induced from a different factor structure, such as bifactor structure and correlated-factor structure.

The remainder of this article is organized as follows: First, the authors review the higher order IRT model. Second, they present the chi-square-based item fit indices for HO-IRT models. Third, the simulation studies are provided to illustrate the performances of those indices. Finally, they end with some concluding remarks.

Method

Model Description

This section introduces the notations and the model descriptions for the higher order IRT models. Interested readers can refer to de la Torre and Song (2009) for a full description.

Usually in the HO-IRT framework, a test is assumed to have a multi-unidimensional structure, in which each item measures one domain-specific ability, and in total, there are T domain-specific abilities. Then an overall ability is extracted from the T domain-specific abilities to explain the common variance among them. The HO-IRT model consists of a measurement model (e.g., the three-parameter logistic model expressed in Equation 1) and a higher order dimensional structure (expressed in Equation 2). Mathematically, the model is specified as

$$p_{ijt} = P\left(y_{ijt} = 1 \middle| a_i, b_i, c_i, \theta_{jt}\right) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(a_i\theta_{jt} - b_i\right)\right]}, \tag{1}$$

$$\theta_{jt} = \lambda_t \xi_j + \varepsilon_{jt}, \tag{2}$$

where y_{ijt} is the response of examinee j ($j = 1, \ldots, N$) on item i ($i = 1, \ldots, I$) which measures domain t ($t = 1, \ldots, T$); ξ_j and θ_{jt} are the overall and tth domain-specific abilities; a_i, b_i, and c_i are the discrimination, threshold, and guessing parameters for item i, respectively; D is a scaling constant, which is set to 1.7; λ_t is the latent regression coefficient of the domain-specific ability θ_{jt}, which derives from the correlation (ρ) between abilities (i.e., $\lambda_t = \sqrt{\rho}$); and lastly, ε_{jt} is the residual term with respect to θ_{jt} conditioning on the overall ability ξ_j. For identification purpose, $\xi_j \sim N(0,1)$ and $\varepsilon_{jt} \sim N(0, 1 - \lambda_t^2)$ so that both domain-specific abilities and overall abilities are put on the same metric (de la Torre & Song, 2009), that is, the standard normal metric.

Chi-Square-Based Item-Fit Indices

According to Hambleton, Swaminathan, and Rogers (1991); Stone (2000); and Orlando and Thissen (2000), a common strategy for assessing item fit of an IRT model can be summarized as follows. (a) Estimate the model parameters (i.e., item parameters, ability parameters, coefficients, and residuals in this study) from a dataset, (b) classify the examinees into K subgroups

according to parameter estimates or test scores, (c) calculate the observed and predicted proportions of correctly/incorrectly responses for each item and each subgroup, and (d) calculate chi-square-based statistics by computing the discrepancy between observed and predicted values. The following four item-fit indices all follow this set of steps for evaluating item-level fit. The difference lies in how the discrepancy is calculated and how examinees are grouped.

The traditional chi-square-based fit indices. Both Q_1 and G^2 are considered to be traditional chi-square-based fit indices. For both of them, examinees are rank-ordered and partitioned into 10 homogeneous subgroups according to their overall ability estimations, as 10 subgroups are sufficient to produce a robust estimation of the discrepancy (Yen, 1981).

Yen's (1981) Q_1 for a dichotomous item i has the form

$$Q_{1i} = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}} + \sum_{k=1}^{10} \frac{N_k[(1 - O_{ik}) - (1 - E_{ik})]^2}{1 - E_{ik}} = \sum_{k=1}^{10} \frac{N_k(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}, \quad (3)$$

where k ($k = 1, \ldots, 10$) represents a homogeneous group of examinees, and N_k is the number of examinees in group k. The observed proportions (O_{ik}) are obtained by calculating the proportion of examinees in group k who answer item i correctly. The expected proportions (E_{ik}) are computed from the model as the mean predicted probability of a correct response in each interval. Yen (1981) showed the degrees of freedom (*df*) associated with Q_1 equaled $10 - m$, where m is the number of model parameters excluded ability parameters. With the higher order IRT model, note that m equals the number of item parameters plus 1. This additional parameter refers to the loading from the second-order factors to the first-order factor; only one loading is considered per item because the item displays the simple structure.

In addition to Q_1, McKinley and Mills (1985) constructed a likelihood ratio G^2 statistics based on computations similar to those used for Q_1. The notations are the same as Equation 3. The correct/incorrect responses for each subgroup are tallied, and G^2 for item i can be computed as

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right) \right]. \quad (4)$$

The *df* associated with G^2 also equals $10 - m$.

Orlando and Thissen's S-X^2 index and S-G^2 index. One known problem with Q_1 and G^2 is that they rely on the estimated ξ for group assignment. Hence, if ξ is not accurately estimated, examinees will likely be misgrouped, leading to ill-behaved fit indices. Orlando and Thissen (2000) proposed *S-X*^{2} and *S-G*^{2} that overcame this limitation. Instead of relying on model-dependent ξ, *S-X*^{2} and *S-G*^{2} rely on test scores (i.e., number-correct [NC] scores). In addition, the expected correct proportions in both *S-X*^{2} and *S-G*^{2} are calculated relying on all response patterns rather than the predicted probability of a correct response.

The key component in *S-X*^{2} and *S-G*^{2} is the expected correct proportion conditioning on different total scores, and Orlando and Thissen's (2000) original idea can be extended to the HO-IRT model and any hierarchical models alike. To be specific, the expected proportion of correct response for item i and kth group has the form

$$E_{ik} = \frac{\iint p_i(\boldsymbol{\theta}) f^{*i}(k - 1|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\xi) \phi(\xi) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\xi}{\iint f(k|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\xi) \phi(\xi) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\xi}, \quad (5)$$

where $\boldsymbol{\theta}$ is the vector of domain-specific abilities, ξ denotes the overall ability, $p_i(\boldsymbol{\theta})$ is the IRF of item i, $f(k|\boldsymbol{\theta})$ is the NC score posterior distribution (Orlando & Thissen, 2000) for score group k, $f^{*i}(k-1|\boldsymbol{\theta})$ is the NC score posterior distribution for score group $k-1$ excluding item i, $p(\boldsymbol{\theta}|\xi)$ is the conditional posterior distribution of domain-specific ability $\boldsymbol{\theta}$ given overall ability ξ, and $\phi(\xi)$ represents the population distribution of ξ. In the original multivariate S-X^2 and S-G^2, $p(\boldsymbol{\theta}|\xi)\phi(\xi)$ is combined as a prior distribution of latent traits, which leads to a T-fold integral. As every item loads onto only one domain-specific dimension, the integral in Equation 5 reduces to a two-fold integral. A rectangular quadrature over equally spaced increments of $\boldsymbol{\theta}$ and ξ from -4.5 to 4.5 (Stroud, 1974) can be used to approximate the integral in Equation 5.

Cai's (2015) Lord–Wingersky algorithm version 2.0 is used to calculate the NC score posterior distributions. The version 2.0 is an extension of Lord and Wingersky's (1984) original algorithm to multiunidimensional structures, that is,

$$f(k|\boldsymbol{\theta}) = \sum_{s_1+\cdots+s_T=k} \prod_{t=1}^{T} L(s_t|\theta_t)q \tag{6}$$

where $L(s_t|\theta_t)$ is the NC score posterior distribution for the items which measure the tth dimension and the NC score is s_t. And $f^{*i}(k-1|\boldsymbol{\theta})$ can be calculated similarly to $f(k|\boldsymbol{\theta})$.

As dichotomous items are focused in this study, the NC score posterior distribution for the tth dimension can be obtained by Lord and Wingersky's (1984) algorithm. For the first item measuring the tth dimension, $L^*(0|\theta_t) = 1 - p_1(1|\theta_t)$ and $L^*(1|\theta_t) = p_1(1|\theta_t)$, where $p_1(1|\theta_t)$ is the probability of a correct response on the first item measured the tth dimension and $L^*(s|\theta_t)$ is the interim value for the NC score posterior distribution for score group s. Then add the second item measuring the tth dimension, $L(0|\theta_t) = L^*(0|\theta_t)(1 - p_2(1|\theta_t))$ and $L(1|\theta_t) = L^*(0|\theta_t)p_2(1|\theta_t) + L^*(1|\theta_t)(1 - p_2(1|\theta_t))$, where $p_2(\theta_t)$ is the IRF of the second item. After adding each item, the new $L(s|\theta_t)$ replaces $L^*(s|\theta_t)$ for all scores computed for the previous item. Repeating this recursive process, the following equation was obtained:

$$L(s_t|\theta_t) = L^*(s_t - 1|\theta_t)p_i(\theta_t) + L^*(s_t|\theta_t)(1 - p_i(\theta_t))q, \tag{7}$$

where $p_i(\theta_t)$ is the IRF of the ith item. Below is a pseudo-algorithm that further details the algorithm.

Cai's Lord–Wingersky Algorithm Version 2.0 for Calculating $f(k|\theta)$

Step 1: Calculate $L(s_t|\theta_t)$ for $s_t = 0, \ldots, I_t$ and $t = 1, \ldots, T$ using Equation 7, where I_t is the total number of items measured the tth dimension.

Step 2: Construct a T-column matrix, $\boldsymbol{S}$, which consisted of all the possible score patterns given total score k. In this matrix, the sum of each row equals the NC score k. The tth column of $\boldsymbol{S}$ consists of all the possible subscores of the tth dimension. Denote S_{ht} as the element in the hth row and the tth column.

Step 3: Calculate $\prod_{t=1}^{T} L(S_{ht}|\theta_t)$ for each h, where $L(S_{ht}|\theta_t)$ is preknown from Step 1, then $f(k|\boldsymbol{\theta})$ can be obtained by summing all of them, as shown in Equation 6.

Given E_{ik} computed from Equation 5, the expressions of S-X^2 and S-G^2 for a dichotomous item i on an I-item test are as follows:

$$S{-}X_i^2 = \sum_{k=1}^{I-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})}, \tag{8}$$

$$S{-}G_i^2 = 2\sum_{k=1}^{I-1} N_k \left[O_{ik} \ln\left(\frac{O_{ik}}{E_{ik}}\right) + (1 - O_{ik}) \ln\left(\frac{1 - O_{ik}}{1 - E_{ik}}\right) \right]. \tag{9}$$

For an I-item test, based on total score, $(I + 1)$ subgroups are naturally formed. At the two extremes, when NC score equals 0 (or I), the proportion of examinees who answered item i correctly is always 0 (or 1). Therefore, the expected probabilities for only $(I - 1)$ subgroups need to be calculated. In Equations 8 and 9, k ($k = 1, \ldots, I - 1$), N_k, and O_{ik} represent the number of examinees in group k, and the observed proportion of item i in group k, respectively, and E_{ik} is the expected proportions of item i in group k. The df associated with S-X^2 and S-G^2 both equals $I - m - 1$, where m is the number of model parameters excluded ability parameters.

Simulation Studies

The main purpose of the simulation studies is to examine the performances of chi-square-based fit indices for HO-IRT models, including the Q_1, G^2, S-X^2, and S-G^2 indices. Two different sources of misfit are considered: The first type of misfit relies on different functional forms assumed for the item characteristic curve (ICC), and the second type of misfit is due to different multivariate latent trait structures, that is, higher order structure, bifactor structure, and simple multidimensional structure. To this end, the simulation studies are divided into two parts, each one focused on one kind of misfit.

Table 1 shows the summary of data generation and calibration models for both studies. To assess the power of indices, the response data are generated from a mixture of two models (denoted as $\text{Model}_A/\text{Model}_B$). More details will be given in the simulation design section.

For all conditions, the Markov chain Monte Carlo (MCMC) algorithm (de la Torre & Hong, 2010; de la Torre & Song, 2009) was used to estimate parameters, and 100 replications were conducted per condition. In each Markov chain, there were 10,000 iterations, half as burn-in phase and half as sampling phase. False positive rates (FPRs) and correct detection rates (CDRs; Wang et al., 2015) were calculated for each condition to evaluate the performances of those indices.

To account for a sampling error associated with expected rejecting rates, as Zhang and Stone (2008) suggested, 95% confidence intervals (CIs) for the true rejection rates were reported:

$$\text{CI}_{95\%} = \alpha \pm 1.96 \times \left[\frac{\alpha(1 - \alpha)}{R}\right]^{1/2},$$

where R is the number of replications (R equals 100 in this study) and α is the significant level (which is set to .05 in this study). Hence, the expected 95% CI is [0.01, 0.09].

Simulation Design 1

Study 1 was designed to examine the performances of chi-square-based fit indices for HO-IRT models to detect the misfit relying on different ICCs. Five factors and their varied conditions were considered: (a) generation model, three different combinations of models with different mixed ratios (i.e., misfit proportions); (b) test structure (de la Torre & Hong, 2010), 40 items measured four dimensions equally, 40 items measured two dimensions equally, or 20 items

Table 1. Summary of Data Generation and Calibration Models.

Study 1						
Calibration model	Generation model					
	1PHO	2PHO	3PHO	2PHO/3PHO	1PHO/3PHO	1PHO/2PHO
1PHO	FPR				FPR&CDR	FPR&CDR
2PHO		FPR		FPR&CDR		
3PHO			FPR			
Study 2						
Calibration model	Generation model					
	2PHO/bifactor			M2PLM		
2PHO	FPR&CDR			FPR		

Note. FPR = false positive rate; CDR = correct detection rate; 1PHO = one-parameter higher order IRT model; 2PHO = two-parameter higher order IRT model; 3PHO = three-parameter higher order IRT model; M2PLM = multidimensional two-parameter logistic model; 2PHO/3PHO = (1 − *proportion*) × 2PHO + *proportion* × 3PHO; 1PHO/3PHO = (1 − *proportion*) × 1PHO + *proportion*× 3PHO; 1PHO/2PHO = (1 − *proportion*) × 1PHO + *proportion*× 2PHO; 2PHO/bifactor= (1 − *proportion*) × 2PHO + *proportion*× bifactor model; 2PHO/M2PLM = (1 − *proportion*) × 2PHO + *proportion*× M2PLM.

measured two dimensions equally; (c) correlation (ρ) between the domains (de la Torre & Hong, 2010), 0.5 (small), 0.7 (medium), or 0.9 (large); (d) sample size (*N*), 1,000 (medium) or 2,000 (large); and (e) misfit proportions (Wang et al., 2015), 0.1 (small), 0.2 (medium), and 0.4 (large), that is, the proportion of misfitting items per domain. The misfitting items were spread equally across multiple dimensions. Totally, there were 162 (3 generation models × 3 test structures × 3 correlations × 2 sample sizes × 3 proportions) different conditions simulated. To generate the response data, the discrimination parameters for each dimension were all distributed from log*N*(0, 0.5), the difficulty parameters were drawn from a standard normal distribution, and the guessing parameters were generated from *Beta*(8, 32). The overall abilities were simulated from a standard normal distribution.

Simulation Results 1

FPRs. Tables 2 and 3 displayed the results of FPRs for different conditions. As the FPRs of both Q_1 and G^2 were much larger than the other two indices, the authors only presented the comparison between $S\text{-}X^2$ and $S\text{-}G^2$, and highlighted the values which exceeded the expected 95% CI; the full results were provided as a supplementary file. $S\text{-}G^2$ had smaller FPRs except for 3PHO/1PHO combination, and for the other conditions, the misfit proportion and the correlation level had nearly no effect, but larger sample size led to higher FPR. For 3PHO/1PHO combination, when the misfit proportion was 40%, $S\text{-}G^2$ had inflated FPR, this inflation would be reduced by small correlation and small sample size. This is because both $S\text{-}X^2$ and $S\text{-}G^2$ are based on total score for grouping; hence, when test length increases, the total number of groups also increases. As a result, the frequency table becomes larger, and the number of small observed/excepted frequencies increases, and small frequency is known to have larger impact on $S\text{-}G^2$ than $S\text{-}X^2$ (Fienberg, 1979), which leads to much higher FPRs of $S\text{-}G^2$ than $S\text{-}X^2$. Furthermore, when I/T (test length/domain) was 40/2 or 20/2, this inflation also appeared for 20% misfit proportion. On the contrary, when the generation model was a single model, $S\text{-}G^2$ also had smaller FPRs,

Table 2. False Positive Rates Under GM and CM identical (GM = CM) in Study 1.

			40/4			40/2			20/2		
N	*p*	Index	1PHO	2PHO	3PHO	1PHO	2PHO	3PHO	1PHO	2PHO	3PHO
1,000	.9	$S\text{-}X^2$	.1073	.0850	.1105	.1080	.0897	.1263	.0914	.0810	.1495
		$S\text{-}G^2$	**.0417**	**.0412**	**.0845**	**.0405**	**.0405**	**.0917**	**.0714**	**.0655**	**.1185**
	.7	$S\text{-}X^2$	.1022	.0910	.1095	.1040	.0910	.1243	.0910	.0940	.1580
		$S\text{-}G^2$	**.0395**	**.0412**	**.0772**	**.0355**	**.0412**	**.0882**	**.0540**	**.0730**	**.1250**
	.5	$S\text{-}X^2$	.1025	.0933	.1235	.1165	.0915	.1515	.1014	.1165	.1570
		$S\text{-}G^2$	**.0403**	**.0420**	**.0955**	**.0393**	**.0362**	**.1077**	**.0640**	**.0760**	**.1280**
2,000	.9	$S\text{-}X^2$	.0985	.0803	.1195	.1028	.0910	.1397	.0880	.0875	.1430
		$S\text{-}G^2$	**.0502**	**.0422**	**.0917**	**.0493**	**.0400**	**.1073**	**.0655**	**.0700**	**.1150**
	.7	$S\text{-}X^2$	.0973	.0945	.1245	.0943	.0973	.1467	.0960	.1030	.1655
		$S\text{-}G^2$	**.0442**	**.0442**	**.0865**	**.0447**	**.0483**	**.1077**	**.0755**	**.0865**	**.1445**
	.5	$S\text{-}X^2$	.0985	.0922	.1230	.1127	.1097	.1472	.0985	.1270	.1650
		$S\text{-}G^2$	**.0367**	**.0445**	**.0943**	**.0522**	**.0510**	**.1065**	**.0760**	**.1010**	**.1505**

Note. The bold values denote the minimal values under each condition. GM = Generation model; CM = Calibration model; 1PHO = one-parameter higher order IRT model; 2PHO = two-parameter higher order IRT model; 3PHO = three-parameter higher order IRT model, CI = confidence interval.

and the performance of $S\text{-}X^2$ was similar to Orlando and Thissen (2000) when I/T was 40/4 and Li and Rupp (2011) when I/T was 40/2. The FPRs for 3PHO were larger than other two models.

Across Tables 2 and 3, $S\text{-}G^2$ had smaller FPRs than $S\text{-}X^2$, but the performance of $S\text{-}X^2$ was more consistent. The influence of sample size on $S\text{-}G^2$ was more notable than that on $S\text{-}X^2$. The FPRs were the largest when I/T was 20/2 and the smallest when I/T was 40/4. In other words, for the same test length, the fewer domain-specific abilities measured, the smaller the FPRs were; when to fix the number of dimensions, larger test length led to smaller FPRs; for the same number of domain-specific items, increasing dimensions would reduce the FPRs. On the whole, $S\text{-}G^2$ performed better based on the results of FPRs because more of $S\text{-}X^2$ conditions exceeded the excepted 95% CI, although $S\text{-}X^2$ performed more consistent across manipulated conditions.

CDRs. Table 4 provided the results of CDRs for different conditions. The CDRs of Q_1 and G^2 were not useful because of their inflated FPRs. Hereafter, the performances of $S\text{-}X^2$ and $S\text{-}G^2$ were compared. As shown in Table 4, lager sample size, higher misfit proportion, and lower correlation all led to higher CDR. For 3PHO/2PHO combination, the CDRs of $S\text{-}X^2$ were larger than those of $S\text{-}G^2$, but the CDRs of both indices were too low to detect misfit. For 3PHO/1PHO condition, $S\text{-}X^2$ also had larger CDR than $S\text{-}G^2$, and the values were the largest among these three combinations. For 2PHO/1PHO combination, $S\text{-}X^2$ and $S\text{-}G^2$ performed similarly to detect misfit.

Comparing the results of all the conditions, the CDR was the largest when I/T was 40/2 and the smallest when I/T was 40/4. It implies that, when to fix the number of dimensions, longer test length leads to higher CDR. In contrast, for the same test length, reducing the number of dimensions helps increase the CDR. Furthermore, if fixing the number of domain-specific items, fewer dimensions led to higher CDR. It appears that chi-square-based indices cannot be used to detect the sole influence of guessing parameter (due to low CDR of differentiating 2PHO vs. 3PHO); nevertheless, they are more sensitive in detecting misfit due to misspecification of discrimination parameters. Without doubt, when the two models differ in both discrimination and guessing parameters, the CDR is the highest.

Table 3. False Positive Rates Under GM More Complex Than CM (GM > CM) in Study 1.

				GM > CM								
				3PHO/2PHO			3PHO/1PHO			2PHO/1PHO		
I/D	*N*	*p*	Index	10%	20%	40%	10%	20%	40%	10%	20%	40%
40/4	1,000	.9	*S-X*^{2}	.0794	.0841	.0896	.0703	.0597	**.0446**	.1014	.1016	.0988
			S-G^2	**.0375**	**.0366**	**.0388**	**.0361**	**.0500**	.1029	**.0439**	**.0397**	**.0450**
		.7	*S-X*^{2}	.0936	.0975	.0883	.0797	.0597	**.0413**	.1053	.1025	.0908
			S-G^2	**.0367**	**.0338**	**.0400**	**.0339**	**.0425**	.0696	**.0358**	**.0425**	**.0292**
		.5	*S-X*^{2}	.0861	.0856	.1008	.0775	.0653	**.0504**	.0975	.0947	.0925
			S-G^2	**.0378**	**.0331**	**.0517**	**.0333**	**.0341**	.0596	**.0303**	**.0338**	**.0400**
	2,000	.9	*S-X*^{2}	.0847	.0850	.0883	.0719	**.0534**	**.0892**	.0967	.1003	.0879
			S-G^2	**.0431**	**.0441**	**.0458**	**.0603**	.0856	.2321	**.0514**	**.0591**	**.0521**
		.7	*S-X*^{2}	.0875	.0906	.1013	.0728	**.0634**	**.0800**	.0922	.0878	.0904
			S-G^2	**.0406**	**.0428**	**.0458**	**.0425**	.0653	.1721	**.0442**	**.0444**	**.0483**
		.5	*S-X*^{2}	.0897	.0906	.0963	.0797	.0609	**.0704**	.0994	.0944	.0887
			S-G^2	**.0436**	**.0437**	**.0471**	**.0436**	**.0478**	.1113	**.0364**	**.0362**	**.0404**
40/4	1,000	.9	*S-X*^{2}	.0881	.0831	.0917	.0683	.0531	**.0454**	.1072	.1103	.1033
			S-G^2	**.0414**	**.0388**	**.0367**	**.0414**	**.0512**	.1017	**.0461**	**.0425**	**.0475**
		.7	*S-X*^{2}	.1017	.0956	.1017	.0811	.0619	**.0517**	.1097	.1179	.0908
			S-G^2	**.0394**	**.0344**	**.0425**	**.0369**	**.0531**	.1058	**.0417**	**.0403**	**.0292**
		.5	*S-X*^{2}	.1056	.1013	.1246	.0786	.0703	**.0487**	.1086	.1069	.1042
			S-G^2	**.0461**	**.0406**	**.0512**	**.0319**	**.0434**	.0696	**.0447**	**.0425**	**.0421**
	2,000	.9	*S-X*^{2}	.0869	.0941	.1025	.0664	**.0622**	**.0908**	.1042	.1044	.0875
			S-G^2	**.0461**	**.0450**	**.0604**	**.0517**	.0953	.2450	**.0611**	**.0625**	**.0579**
		.7	*S-X*^{2}	.0917	.1084	.1088	.0750	**.0625**	**.0858**	.1017	.0953	.0983
			S-G^2	**.0461**	**.0525**	**.0629**	**.0586**	.0794	.1979	**.0561**	**.0547**	**.0600**
		.5	*S-X*^{2}	.1092	.1088	.1213	.0819	**.0728**	**.0867**	.1244	.0988	.0975
			S-G^2	**.0539**	**.0525**	**.0642**	**.0533**	.0744	.1713	**.0553**	**.0462**	**.0554**
20/2	1,000	.9	*S-X*^{2}	.0933	.0869	.0867	.0761	**.0656**	**.0917**	.1022	.1044	.0958
			S-G^2	**.0694**	**.0537**	**.0575**	**.0750**	.1019	.2075	**.0689**	**.0894**	**.0800**
		.7	*S-X*^{2}	.0939	.1025	.0950	.0817	**.0706**	**.0800**	.0789	.0869	.1108
			S-G^2	**.0767**	**.0719**	**.0592**	**.0789**	.0925	.1800	**.0617**	**.0663**	**.0900**
		.5	*S-X*^{2}	.1039	.1138	.1058	.0778	**.0781**	**.0817**	.1078	.1025	.1158
			S-G^2	**.0750**	**.0831**	**.0767**	**.0700**	.0856	.1467	**.0750**	**.0681**	**.0867**
	2,000	.9	*S-X*^{2}	.1078	.1031	.1108	**.0856**	**.1163**	**.2625**	.0917	.1056	**.1300**
			S-G^2	**.0750**	**.0819**	**.0967**	.1039	.1719	.3858	**.0772**	**.1019**	.1308
		.7	*S-X*^{2}	.1117	.1019	.1333	**.0917**	**.0969**	**.2292**	.0917	.1281	**.1125**
			S-G^2	**.0906**	**.0719**	**.1092**	.1011	.1644	.3642	**.0828**	**.1000**	.1133
		.5	*S-X*^{2}	.1156	.1225	.1283	**.0872**	**.1100**	**.2125**	.0950	.1100	**.1325**
			S-G^2	**.0844**	**.0906**	**.0992**	.0950	.1300	.3150	**.0861**	**.0981**	.1333

Note. The bold values denote the minimal values under each condition. GM = Generation model; CM = Calibration model; 3PHO = three-parameter higher order IRT model; 2PHO = two-parameter higher order IRT model; 1PHO = one-parameter higher order IRT model; CI = confidence interval.

Simulation Design 2

Study 2 was designed to further investigate the performances of *S-X*^{2} and *S-G*^{2}, which were considered acceptable in Study 1, in the context of detecting items conforming to a different factor structure. Table 5 showed the different factor structures that were considered in this study. The HO-IRT model is a special version of the bifactor model, which adds the proportionality constraints on the general factor and group factor discrimination parameters for each domain, so that the HO-IRT model is nested within the bifactor model. Also, the HO-IRT model can be

Table 4. Correct Detection Rates Under GM More Complex Than CM (GM > CM) in Study 1.

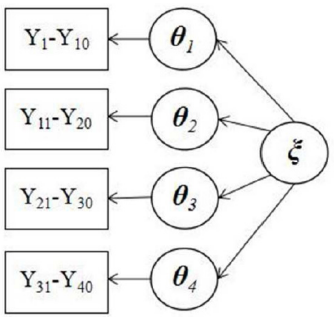
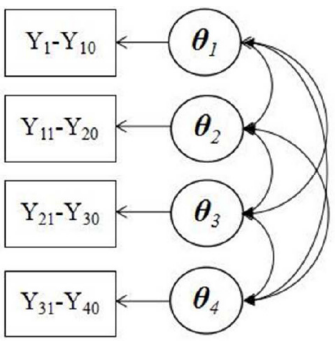
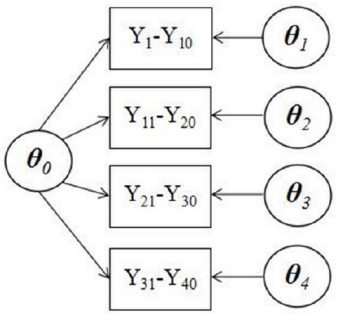
				3PHO/2PHO			3PHO/1PHO			2PHO/1PHO		
I/D	*N*	*p*	Index	10%	20%	40%	10%	20%	40%	10%	20%	40%
40/4	1,000	.9	S-X^2	.2500	.2587	.2094	**.7750**	**.6700**	**.6169**	**.4400**	**.4838**	.4531
			S-G^2	.1950	.1750	.1494	.7225	.6125	.5919	**.4400**	.4700	.4625
		.7	S-X^2	.1850	.1600	.1600	.6375	.5587	**.5331**	.4250	.3787	.3669
			S-G^2	.1300	.0925	.0988	.5750	.4475	.4931	.4200	.3950	.3750
		.5	S-X^2	.1575	.1350	.1269	**.6200**	**.5475**	**.4456**	.3250	.2913	.3187
			S-G^2	.0825	.0862	.0819	.5100	.4188	.3794	.2825	.3013	.3088
	2,000	.9	S-X^2	.4025	.3575	.2775	.8250	.7863	.7512	.5850	.6088	.6062
			S-G^2	.3500	.3125	.2200	.8150	.7750	.7406	**.6050**	.6150	.6150
		.7	S-X^2	.2875	.2562	.2025	.7375	.7163	**.6719**	**.5700**	.5363	.5594
			S-G^2	.2700	.1888	.1581	.7150	.6613	.6462	.5650	.5400	.5594
		.5	S-X^2	.1675	.1812	.1812	.6900	.6288	.5837	.4225	.4338	.4537
			S-G^2	.1050	.1300	.1138	.6025	.5863	.5706	.4450	.4238	.4400
40/2	1,000	.9	S-X^2	.2650	.2488	.2238	**.7425**	**.7100**	**.6194**	**.5175**	.4875	**.4894**
			S-G^2	.1925	.1950	.1525	.6600	.6613	.5837	.4825	**.5038**	**.4894**
		.7	S-X^2	.2525	.2188	.1800	**.7100**	**.6538**	**.5813**	.4275	.4400	.3669
			S-G^2	.1925	.1375	.1156	.6225	.5813	.5319	.4600	.4350	.3750
		.5	S-X^2	.1775	.1863	.1700	**.6425**	.6275	.5081	.3550	.3912	.3688
			S-G^2	.1300	.1075	.1019	.5500	.5525	.4631	.3150	.3937	.3569
	2,000	.9	S-X^2	**.3850**	.3675	.3113	**.8475**	**.8063**	**.7712**	.6175	.6375	.6125
			S-G^2	.3225	.3225	.2575	.8225	.7837	.7581	**.6325**	**.6600**	**.6219**
		.7	S-X^2	.3175	.3250	.2581	**.8200**	**.7588**	.7244	.5475	.5713	.5869
			S-G^2	.2450	.2675	.2050	.7800	.7087	.6950	.5275	.6012	.6094
		.5	S-X^2	.2800	.2600	.1944	.7650	**.7300**	.6675	.5350	.5138	.5406
			S-G^2	.2175	.1913	.1400	.7475	.7025	.6525	.5150	.5050	.5419
20/2	1000	.9	S-X^2	.3200	.2425	.2188	**.7500**	**.7675**	.6875	.5300	.5475	.5550
			S-G^2	.2950	.2125	.2062	.7200	.7400	.6637	.5450	.5625	.5463
		.7	S-X^2	.2550	.2125	.2037	**.7400**	**.6775**	.6200	.5500	.5200	.5313
			S-G^2	.2250	.1800	.1737	.6350	.6625	.5888	.5550	.5225	.5425
		.5	S-X^2	.1850	.1450	.1750	.5950	.6325	.5663	.4850	.4550	.4500
			S-G^2	.1600	.1125	.1425	.5250	.6000	.5700	.4950	.4875	.4663
	2,000	.9	S-X^2	.3950	.3525	.2950	.8350	**.8200**	.7837	.6150	**.6925**	.6725
			S-G^2	.3700	.3550	.2900	**.8450**	.8025	.7950	**.6300**	.6750	.6863
		.7	S-X^2	.3600	.3000	.2600	**.7700**	**.8275**	.7588	.6150	.6200	.6550
			S-G^2	.3450	.2625	.2425	.7550	.8225	.7800	.6500	.6500	.6713
		.5	S-X^2	.2800	.2375	.2200	.8050	.7375	.6900	.6150	.6575	.6025
			S-G^2	.2350	.2300	.2025	.7850	.7300	.6937	.6100	.6650	.6050

Note. The bold values denotes the maximum values under each condition. GM = Generation model; CM = Calibration model; 3PHO = three-parameter higher order IRT model; 2PHO = two-parameter higher order IRT model; 1PHO = one-parameter higher order IRT model.

considered nested within the simple-structure MIRT model when the number of dimensions, D, exceeds 3; when $D = 3$, the two models will give exactly the same fit.

According to the results of Study 1, only one level sample size (i.e., $N = 1{,}000$) and one level test structure (i.e., 40 items measured four dimensions equally) were considered. There were 10 (1 MIRT condition + 9 bifactor conditions: 3 correlations × 3 proportions) different conditions simulated. Because the difference between MIRT and the HO-IRT models only emerges when we look at the factor covariance matrix, and item parameters of both models have the same meanings. Hence, if the true data were generated from the MIRT model and retrofitted with the HO-IRT model, one would expect the detection rate of $S\text{-}X^2$ and $S\text{-}G^2$ be close to the nominal

Table 5. Summary of Comparison Among Different Multidimensional Structures.

	HO-IRT	MIRT	Bifactor
Test structure			
Formulation	$P(y_{ijt}=1\mid a_i,b_i,\theta_{jt})=\frac{1}{1+\exp\left[-D\left(a_i\theta_{jt}-b_i\right)\right]}$ $\theta_{jt}=\lambda_t\xi_j+\varepsilon_{jt}$	$P(y_{ijt}=1\mid a_i,b_i,\theta_{jt})=\frac{1}{1+\exp\left[-D\left(a_i\theta_{jt}-b_i\right)\right]}$	$P(y_{ijt}=1\mid a_{i0},a_{it},b_i,\theta_{j0},\theta_{jt})=\frac{1}{1+\exp\left[-D\left(a_{i0}\theta_{j0}+a_{it}\theta_{jt}-b_i\right)\right]}$
Latent traits structure	$\begin{bmatrix}\xi\\ \theta_1\\ \vdots\\ \theta_4\end{bmatrix}\sim MVN\left(\begin{bmatrix}0\\0\\ \vdots\\0\end{bmatrix}\begin{bmatrix}1 & \lambda_1^2 & \cdots & \lambda_4^2\\ \lambda_1^2 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ \lambda_4^2 & 0 & \cdots & 1\end{bmatrix}\right)$	$\begin{bmatrix}\theta_1\\ \vdots\\ \theta_4\end{bmatrix}\sim MVN\left(\begin{bmatrix}0\\ \vdots\\0\end{bmatrix}\begin{bmatrix}1 & \cdots & \rho_{14}\\ \vdots & \ddots & \vdots\\ \rho_{14} & \cdots & 1\end{bmatrix}\right)$	$\begin{bmatrix}\theta_0\\ \theta_1\\ \vdots\\ \theta_4\end{bmatrix}\sim MVN\left(\begin{bmatrix}0\\0\\ \vdots\\0\end{bmatrix}\begin{bmatrix}1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1\end{bmatrix}\right)$
Notation	a_i, discrimination parameter b_i, difficulty parameter θ_{jt}, domain-specific ability ξ_j, overall ability ε_{jt}, residual D, a scaling constant ($i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T$)	a_i, discrimination parameter b_i, difficulty parameter θ_{jt}, ability D, a scaling constant ($i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T$)	a_{i0}, general discrimination parameter a_{it}, group discrimination parameter b_i, difficulty parameter θ_{j0}, general ability θ_{jt}, group ability D, a scaling constant ($i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T$)
#Parameter	$2\times I+T$	$2\times I+T\times(T-1)/2$	$3\times I$

Note. Take four-dimension structure as an example. HO-IRT = higher order item response theory; MIRT = multidimensional item response theory.

Table 6. Comparison Between $S\text{-}X^2$ and $S\text{-}G^2$ When the Misspecifying Items Have a Bifactor Structure.

		10%		20%		40%	
ρ	Index	FPR	CDR	FPR	CDR	FPR	CDR
.9	$S\text{-}X^2$	.0847	.1100	.0894	.0825	.0950	.0869
	$S\text{-}G^2$	.0397	.0525	.0453	.0362	.0454	.0400
.7	$S\text{-}X^2$	.0919	.0700	.1016	.1225	.1021	.1256
	$S\text{-}G^2$	.0436	.0300	.0494	.0650	.0471	.0737
.5	$S\text{-}X^2$	.0942	.1400	.1166	.1375	.1225	.1431
	$S\text{-}G^2$	.0400	.1600	.0537	.1388	.0579	.1275

Note. FPR = false positive rate; CDR = correct detection rate.

level. Therefore, in this case, the authors do not manipulate the proportion of misfit, but rather assume all items conform to the MIRT model. This provides an additional sanity check on the proposed indices. To generate the response data using the MIRT model, the discrimination parameters were all distributed from log$N$(0, 0.5), the difficulty parameters were drawn from a standard normal distribution, and the correlations between two different dimensions were all set to 0.5.

On the contrary, the difference between the bifactor and HO-IRT models could show up at item level. Therefore, when the misfitting item had a bifactor structure, the mixed models with three different mixed ratios were considered as the data generation model. The selections of correlations (ρ) and misfit proportions were the same as those in Study 1. To generate the response data, the discrimination parameters were all distributed from log$N$(0, 0.5), and the difficulty parameters were all drawn from a standard normal distribution. And for the misfitting items conforming to the bifactor structure, the corresponding group discrimination parameter was regenerated from log$N$(0, 0.5), and the general discrimination parameter was the same as the original discrimination parameters in the HO-IRT model. The abilities for each domain were all generated from a standard normal distribution.

Simulation Results 2

When the misfitting items were generated from the bifactor model (Table 6), $S\text{-}G^2$ had smaller FPRs than $S\text{-}X^2$ under all conditions, which was consistent with Study 1, but the manipulated factors did not have a consistent effect on the FPRs. Regarding CDR, $S\text{-}X^2$ had smaller CDRs than $S\text{-}G^2$ except for the condition with medium to large correlation and high misfit proportion. As one would expect, higher correlation makes both the bifactor model and the HO-IRT models close to a UIRT model, and the distinction between them becomes so small that the CDR is low. Higher misfit proportion also led to smaller CDR. One explanation is, when there is a large proportion of misfitting items, the item parameter estimates would likely be biased due to the contamination of the misfitting items, and therefore, the item-level misfit detection becomes more difficult. In fact, this is known as the disadvantage of almost all residual-based (or discrepancy-based) fit indices. CDR was relatively low across all conditions with values ranging from .083 to .143 for $S\text{-}X^2$ and from .030 to .160 for $S\text{-}G^2$. This is not surprising because when fitting the HO-IRT model to data simulated from a combination of the HO-IRT and the bifactor model, recovery of the entire test is acceptable, despite the misfitting items cannot be recovered well. Indeed, the observed results further support that these indices would not have inflated FPRs.

As a reference, when the response data were generated solely from the bifactor model and fitted with the HO-IRT model (i.e., the misfit proportion was 100%), the CDRs of $S\text{-}X^2$ and $S\text{-}G^2$ were .285 and .257, respectively. These values are higher than those reported in Li and Rupp (2011), in which they tried to detect item-level misfit when data were generated from the bifactor model and fitted with the MIRT model.

When the response data were generated from the MIRT model and the HO-IRT model was used to fit the data (i.e., the misfit proportion was 100%), the CDRs of $S\text{-}X^2$ and $S\text{-}G^2$ were .147 and .098, respectively, which were consistent with Li and Rupp (2011). This observed low power is not unexpected, however, because the parameter estimates from the ''misfitting'' HO-IRT model were actually close to the true MIRT model parameters. In particular, the root mean square error between the estimated and true domain-specific abilities was in the range of .448 to .498, and the bias was in the range of –.026 to –.009. Item parameter recovery was also acceptable. Due to the close resemblance between the HO-IRT and MIRT models, the low CDR is actually reassuring because it implies the FPR is well controlled.

Discussion

Before any model-based inferences can be drawn, the model's fit must be thoroughly assessed, because any conclusion derived from poorly fitting models may be potentially misleading. In practice, the item-fit will not be analyzed solely. Actually, the model-data fit at the global model level must be investigated firstly using model fit indices, when a model does not fit well, alternative models might be fitted. However, more often than not, no such model provides a good fit (Liu & Maydeu-Olivares, 2014). Facing this situation, researchers have to differentiate well-fitting items from poorly fitting ones; then they may decide to retain only the well-fitting set or to apply an alternative IRT model to the poorly fitting set on the basis of item fit analysis. In other words, item-level fit analysis not only serves as a complementary check to global fit analysis, it is also essential in scale development because the fit result will help guide item revision or deletion (Liu & Maydeu-Olivares, 2014). Although there are abundant research focusing on item-level fit evaluation for both unidimensional and multidimensional IRT models, there is lack of an effective item-level fit index for hierarchical models, and the aim of the present study is to fill in this gap. Moreover, there is also not enough information on which version of chi-square-based item fit indices is recommended for HO-IRT models under different conditions. Hence, another main purpose of this study was to compare the performances of chi-square-based item fit indices for HO-IRT models. Last but not least, there is rare research on how item fit indices perform to detect the misfit relying on different latent trait structure; the other goal of this study was to examine the power of item fit indices to compare among HO-IRT models, bifactor models, and MIRT models.

Across all simulation conditions, $S\text{-}G^2$ is recommended for HO-IRT models due to its smaller FPR and adequate CDR. Similar to the findings reported in the literature, both $S\text{-}X^2$ and $S\text{-}G^2$ perform poorly to detect model misspecification due to guessing behavior at the item level and perform well to detect different discrimination scales. Furthermore, both $S\text{-}X^2$ and $S\text{-}G^2$ perform too poorly to detect item misspecification due to multivariate structure of latent traits.

Because $S\text{-}X^2$ and $S\text{-}G^2$ performed well in this study, when the source of misfit was the inaccurate functional form assumed for the ICC, it may be useful to extend the procedures to other hierarchical models such as third-order IRT models (Rijmen, Jeon, von Davier, & Rabe-Hesketh, 2014), response-time models (van der Linden, 2009), and multilevel IRT models (Fox & Glas, 2001). Also, it is easy to extend the study to deal with polytomous data. On the contrary, as Li and Rupp (2011) compared the performances of model fit indices to detect the misfit at the global model level, the performances of model fit indices should be further compared

to detect the misfit at the local item level. In addition, it is not rare for latent traits to distribute nonnormally such as in personality or psychopathology measures (Micceri, 1989); the consequences of normality violation on item-level fit should be further assessed. Finally, it would be worthwhile to compare chi-square-based indices with other fit statistics (Douglas & Cohen, 2001; Glas & Suárez-Falcón, 2003; Haberman, 2009; Haberman et al., 2013; Sinharay, 2006) for hierarchical models.

Acknowledgments

The authors thank the Editor in Chief Dr. Hua-Hua Chang, the Associate Editor Dr. Daniel Bolt and two anonymous reviewers for their helpful comments on earlier drafts of this article.

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 research was supported by the National Natural Science and Social Science Foundations of China (Grant 11571069) and Institute of Education Sciences (IES) (Grant R305D160010).

Supplemental Material

Supplemental material is available for this article online.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed., Revised and expanded). New York, NY: Marcel Dekker.

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.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.

Bock, R. D., & Haberman, S. J. (2009, July). *Confidence bands for examining goodness-of-fit of estimated item response functions*. Paper presented at Annual Meeting of the Psychometric Society, Cambridge, UK.

Cai, L. (2015). Lord-Wingersky algorithm version 2.0 for hierarchical item factor models with applications in test scoring, scale alignment, and model fit testing. *Psychometrika*, *80*, 535-559.

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.

Chon, K. H., Lee, W. C., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*, 318-338.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267-285.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*, 620-639.

Demars, C. E. (2005). Type I error rates for Parscale's fit index. *Educational and Psychological Measurement*, *65*, 42-50.

Douglas, J., & Cohen, A. S. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*, 234-243.

Fienberg, S. E. (1979). The use of chi-squared statistics for categorical data problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, *41*, 54-64.

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

Glas, C. A. W., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement*, *27*, 87-106.

Haberman, S. J. (2009). Use of generalized residuals to examine goodness of fit of item response models. *ETS Research Report Series*, 2009(1), 1-17.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika*, *78*, 417-440.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Huang, H. Y. (2015). A multilevel higher order item response theory model for measuring latent growth in longitudinal data. *Applied Psychological Measurement*, *39*, 362-372.

Huang, H. Y. (2017). Mixture IRT model with a higher-order structure for latent traits. *Educational and Psychological Measurement*, *77*, 275-304.

Huang, H. Y., Chen, P. H., & Wang, W. C. (2012). Computerized adaptive testing using a class of high-order item response theory models. *Applied Psychological Measurement*, *36*, 689-706.

Huang, H. Y., & Wang, W. C. (2013). Higher order testlet response models for hierarchical latent traits and testlet-based items. *Educational and Psychological Measurement*, *73*, 491-511.

Huang, H. Y., Wang, W. C., Chen, P. H., & Su, C. M. (2013). Higher-order item response models for hierarchical latent traits. *Applied Psychological Measurement*, *37*, 619-637.

Huo, Y., de la Torre, J., Mun, E. Y., Kim, S. Y., Ray, A. E., Jiao, Y., & White, H. R. (2015). A hierarchical multi-unidimensional IRT approach for analyzing sparse, multi-group data for integrative data analysis. *Psychometrika*, *80*, 834-855.

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory item fit indices for the graded response model. *Organizational Research Methods*, *14*, 10-23.

Lee, M. (2014). *Application of higher-order IRT models and hierarchical IRT models to computerized adaptive testing* (Electronic Theses and Dissertations). University of California, Los Angeles.

Li, Y., & Rupp, A. A. (2011). Performance of the S-X2 statistic for full-information bifactor models. *Educational and Psychological Measurement*, *71*, 986-1005.

Liang, T., & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement*, *69*, 913-928.

Liang, T., & Wells, C. S. (2015). A nonparametric approach for assessing goodness-of-fit of IRT models in a mixed format test. *Applied Measurement in Education*, *28*, 115-129.

Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An assessment of the nonparametric approach for evaluating the fit of item response models. *Journal of Educational Measurement*, *51*, 1-17.

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

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

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score ''equatings.'' *Applied Psychological Measurement*, *8*, 453-461.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S-X^2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*, 391-406.

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.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49-57.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Muraki, E., & Bock, R. D. (1997). *PARSCALE: IRT item analysis and test scoring for rating-scale data*. [Computer software]. Chicago, IL: Scientific Software International.

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.

Ranger, J., & Kuhn, J. T. (2012). Assessing fit of item response models using the information matrix test. *Journal of Educational Measurement*, *49*, 247-268.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York, NY: Springer.

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, *61*, 331-360.

Rijmen, F., Jeon, M., von Davier, M., & Rabe-Hesketh, S. (2014). A third-order item response theory model for modeling the effects of domains and subdomains in large-scale educational assessment surveys. *Journal of Educational and Behavioral Statistics*, *39*, 235-256.

Roberts, J. S. (2008). Modified likelihood-based item fit statistics for the generalized graded unfolding model. *Applied Psychological Measurement*, *32*, 407-423.

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

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British Journal of Mathematical and Statistical Psychology*, *59*, 429-449.

Stone, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, *37*, 158-175.

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement*, *40*, 331-352.

Stroud, A. H. (1974). *Numerical quadrature and solution of ordinary differential equations*. New York, NY: Springer.

Suárez-Falcón, J. C., & Glas, C. A. W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology*, *56*, 127-143.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, *46*, 247-272.

Wang, C. (2014). Improving measurement precision of hierarchical latent traits using adaptive testing. *Journal of Educational and Behavioral Statistics*, *39*, 452-477.

Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 455-465.

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement*, *39*, 525-538.

Wells, C. S., & Bolt, D. M. (2008). Investigation of a nonparametric procedure for assessing goodness-of-fit in item response theory. *Applied Measurement in Education*, *21*, 22-40.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

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