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).

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

Xue Zhang¹, Chun Wang² and Jian Tao¹

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 $S-X^2$, and $S-G^2$) 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 $S-G^2$ is promising for higher order items.

Keywords
higher order IRT models, item fit, $S-X^2$, $S-G^2$, 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,
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., $Q_1$, $G^2$, $S-X^2$, and $S-G^2$) 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). $S-X^2$ 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 $\chi^2$ and $G^2$ were used to detect item parameter drift (LaHuis et al., 2011; Stone & Zhang, 2003). Also, $S-X^2$ 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 PPCM method framework, such as Yen’s $Q_1$ (Sinharay, 2005; Wang et al., 2015), Stone’s $\chi^2$ (Wang et al., 2015), and Orlando and Thissen’s $S-X^2$ and $S-G^2$ (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 $S-X^2$ 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 $S-X^2$ 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 $S-X^2$ with more than two dimensions was not examined; (b) the performance of $S-X^2$ for comparing the bifactor model with the HO-IRT model was not mentioned; and (c) the recursive algorithm which was needed to compute $S-X^2$ 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: $Q_1$, $G^2$, $S-X^2$, and $S-G^2$ 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

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

$$\theta_{jt} = \lambda_t \xi_j + \epsilon_{jt},$$

where $y_{ijt}$ is the response of examinee $j$ ($j = 1, \ldots, N$) on item $i$ ($i = 1, \ldots, I$) which measures domain $t$ ($t = 1, \ldots, T$); $\xi_j$ and $\theta_{jt}$ are the overall and $t$th 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; $\lambda_t$ is the latent regression coefficient of the domain-specific ability $\theta_{jt}$, which derives from the correlation ($\rho$) between abilities (i.e., $\lambda_t = \sqrt{\rho}$); and lastly, $\epsilon_{jt}$ is the residual term with respect to $\theta_{jt}$ conditioning on the overall ability $\xi_j$. For identification purpose, $\xi_j \sim N(0,1)$ and $\epsilon_{jt} \sim N(0, 1 - \lambda_t^2)$ 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

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

(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

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

(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 \( \xi \) for group assignment. Hence, if \( \xi \) 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 \( \xi \), \( S-X^2 \) 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 \( k \)th group has the form

\[
E_{ik} = \frac{\int p_i(0) f^{\ast i}(k - 1|0)p(0|\xi) \phi(\xi) d\theta d\xi}{\int f(k|0)p(0|\xi) \phi(\xi) d\theta d\xi},
\]

(5)
where $\mathbf{\theta}$ is the vector of domain-specific abilities, $\xi$ denotes the overall ability, $p_i(\mathbf{\theta})$ is the IRF of item $i$, $f(k|\mathbf{\theta})$ is the NC score posterior distribution (Orlando & Thissen, 2000) for score group $k$, $f^{\text{f}}(k−1|\mathbf{\theta})$ is the NC score posterior distribution for score group $k−1$ excluding item $i$, $p(\mathbf{\theta}|\xi)$ is the conditional posterior distribution of domain-specific ability $\mathbf{\theta}$ given overall ability $\xi$, and $\phi(\xi)$ represents the population distribution of $\xi$. In the original multivariate $S-\chi^2$ and $S-G^2$, $p(\mathbf{\theta}|\xi)\phi(\xi)$ 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 $\mathbf{\theta}$ and $\xi$ 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,

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

where $L(s_t|\theta_t)$ is the NC score posterior distribution for the items which measure the $t$th dimension and the NC score is $s_t$. And $f^{\text{f}}(k−1|\mathbf{\theta})$ can be calculated similarly to $f(k|\mathbf{\theta})$.

As dichotomous items are focused in this study, the NC score posterior distribution for the $t$th dimension can be obtained by Lord and Wingersky’s (1984) algorithm. For the first item measuring the $t$th dimension, $L^*(0|\theta_t) = 1 - p_{1t}(1|\theta_t)$ and $L^*(1|\theta_t) = p_{1t}(1|\theta_t)$, where $p_{1t}(1|\theta_t)$ is the probability of a correct response on the first item measured the $t$th 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 $t$th dimension, $L(0|\theta_t) = L^*(0|\theta_t)(1 - p_{2t}(1|\theta_t))$ and $L(1|\theta_t) = L^*(0|\theta_t)p_{2t}(1|\theta_t) + L^*(1|\theta_t)(1 - p_{2t}(1|\theta_t))$, where $p_{2t}(\theta_t)$ 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:

$$L(s_t|\theta_t) = L^*(s_t - 1|\theta_t)p_{1t}(\theta_t) + L^*(s_t|\theta_t)(1 - p_{1t}(\theta_t))q,$$

where $p_{1t}(\theta_t)$ is the IRF of the $t$th item. Below is a pseudo-algorithm that further details the algorithm.

**Cai’s Lord–Wingersky Algorithm Version 2.0 for Calculating $f(k|\mathbf{\theta})$**

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 $t$th dimension.

Step 2: Construct a $T$-column matrix, $\mathbf{S}$, 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 $t$th column of $\mathbf{S}$ consists of all the possible subscores of the $t$th dimension. Denote $S_{ht}$ as the element in the $h$th row and the $t$th 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 $f(k|\mathbf{\theta})$ can be obtained by summing all of them, as shown in Equation 6.

Given $E_{ik}$ computed from Equation 5, the expressions of $S-\chi^2$ and $S-G^2$ for a dichotomous item $i$ on an $I$-item test are as follows:
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 1), 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 Model A/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:

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

where $R$ is the number of replications ($R$ equals 100 in this study) and $\alpha$ 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...
measured two dimensions equally; (c) correlation \( r \) 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 \( \times 3 \) test structures \( \times 3 \) correlations \( \times 2 \) sample sizes \( \times 3 \) proportions) different conditions simulated.

To generate the response data, the discrimination parameters for each dimension were all distributed from \( \log \text{N}(0, 0.5) \), the difficulty parameters were drawn from a standard normal distribution, and the guessing parameters were generated from \( \text{Beta}(8, 32) \). 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 \( S-X^2 \) and \( S-G^2 \), and highlighted the values which exceeded the expected 95% CI; the full results were provided as a supplementary file. \( S-G^2 \) 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%, \( S-G^2 \) had inflated FPR, this inflation would be reduced by small correlation and small sample size. This is because both \( S-X^2 \) and \( S-G^2 \) 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/expected frequencies increases, and small frequency is known to have larger impact on \( S-G^2 \) than \( S-X^2 \) (Fienberg, 1979), which leads to much higher FPRs of \( S-G^2 \) than \( S-X^2 \). 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, \( S-G^2 \) also had smaller FPRs.

### Table 1. Summary of Data Generation and Calibration Models.

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Calibration model</th>
<th>Generation model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1PHO</td>
<td>2PHO</td>
</tr>
<tr>
<td>FPR</td>
<td>FPR&amp;CDR</td>
<td>FPR&amp;CDR</td>
</tr>
<tr>
<td>Study 2</td>
<td>Calibration model</td>
<td>Generation model</td>
</tr>
<tr>
<td></td>
<td>2PHO/bifactor</td>
<td>M2PLM</td>
</tr>
<tr>
<td>FPR &amp;CDR</td>
<td>FPR</td>
<td></td>
</tr>
</tbody>
</table>

*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 = \( (1 - \text{proportion}) \times 2PHO + \text{proportion} \times 3PHO \); 1PHO/3PHO = \( (1 - \text{proportion}) \times 1PHO + \text{proportion} \times 3PHO \); 1PHO/2PHO = \( (1 - \text{proportion}) \times 1PHO + \text{proportion} \times 2PHO \).
and the performance of $S-\chi^2$ 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, $S-G^2$ had smaller FPRs than $S-\chi^2$, but the performance of $S-\chi^2$ was more consistent. The influence of sample size on $S-G^2$ was more notable than that on $S-\chi^2$. 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, $S-G^2$ performed better based on the results of FPRs because more of $S-\chi^2$ conditions exceeded the excepted 95% CI, although $S-\chi^2$ 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 $S-\chi^2$ and $S-G^2$ were compared. As shown in Table 4, larger sample size, higher misfit proportion, and lower correlation all led to higher CDR. For 3PHO/2PHO combination, the CDRs of $S-\chi^2$ were larger than those of $S-G^2$, but the CDRs of both indices were too low to detect misfit. For 3PHO/1PHO condition, $S-\chi^2$ also had larger CDR than $S-G^2$, and the values were the largest among these three combinations. For 2PHO/1PHO combination, $S-\chi^2$ and $S-G^2$ 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 2. False Positive Rates Under GM and CM identical (GM = CM) in Study 1.

<table>
<thead>
<tr>
<th>N</th>
<th>$p$</th>
<th>Index</th>
<th>40/4</th>
<th></th>
<th>40/2</th>
<th></th>
<th>20/2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1PHO</td>
<td>2PHO</td>
<td>3PHO</td>
<td>1PHO</td>
<td>2PHO</td>
<td>3PHO</td>
</tr>
<tr>
<td>1,000</td>
<td></td>
<td>$S-\chi^2$</td>
<td>.1073</td>
<td>.0850</td>
<td>.1105</td>
<td>.1080</td>
<td>.0897</td>
<td>.1263</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.0417</td>
<td>.0412</td>
<td>.0845</td>
<td>.0405</td>
<td>.0405</td>
<td>.0917</td>
</tr>
<tr>
<td>2,000</td>
<td></td>
<td>$S-\chi^2$</td>
<td>.1022</td>
<td>.0910</td>
<td>.1095</td>
<td>.1040</td>
<td>.0910</td>
<td>.1243</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.0395</td>
<td>.0412</td>
<td>.0772</td>
<td>.0355</td>
<td>.0412</td>
<td>.0882</td>
</tr>
</tbody>
</table>

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.
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

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
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 $S-X^2$ and $S-G^2$ be close to the nominal

<table>
<thead>
<tr>
<th>l/D</th>
<th>$N$</th>
<th>$p$</th>
<th>Index</th>
<th>10% 20% 40%</th>
<th>10% 20% 40%</th>
<th>10% 20% 40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>40/4 1,000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.2500 .2587 .2094</td>
<td>.7750 .6700 .6169</td>
<td>.4400 .4838 .4531</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1950 .1750 .1494</td>
<td>.7225 .6125 .5919</td>
<td>.4400 .4700 .4625</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.1850 .1600 .1600</td>
<td>.6375 .5587 .5331</td>
<td>.4250 .3787 .3669</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1300 .0925 .0988</td>
<td>.5750 .4475 .4931</td>
<td>.4200 .3950 .3750</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.1575 .1350 .1269</td>
<td>.6200 .5475 .4456</td>
<td>.3250 .2913 .3187</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.0825 .0862 .0819</td>
<td>.5100 .4188 .3794</td>
<td>.2825 .3013 .3088</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.4025 .3575 .2775</td>
<td>.8250 .7863 .7512</td>
<td>.5850 .6088 .6062</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.3500 .3125 .2200</td>
<td>.8150 .7750 .7406</td>
<td>.6050 .6150 .6150</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.2875 .2562 .2025</td>
<td>.7375 .7163 .6719</td>
<td>.5700 .5363 .5594</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2700 .1888 .1581</td>
<td>.7150 .6613 .6462</td>
<td>.5650 .5400 .5594</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.1675 .1812 .1812</td>
<td>.6900 .6288 .5837</td>
<td>.4225 .4338 .4537</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1050</td>
<td>.5081</td>
<td>.4450 .4238 .4400</td>
<td></td>
</tr>
<tr>
<td>40/2 1,000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.2650 .2488 .2238</td>
<td>.7425 .7100 .6194</td>
<td>.5175 .4875 .4894</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1925 .1950 .1525</td>
<td>.6600 .6613 .5837</td>
<td>.4825 .5038 .4894</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.2525 .2188 .1800</td>
<td>.7100 .6538 .5813</td>
<td>.4275 .4400 .3669</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1925 .1375 .1156</td>
<td>.6225 .5813 .5319</td>
<td>.4600 .4350 .3750</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.1775 .1863 .1700</td>
<td>.6425 .6275 .5081</td>
<td>.3550 .3912 .3688</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1300</td>
<td>.5081</td>
<td>.4450 .4238 .4400</td>
<td></td>
</tr>
<tr>
<td>2,000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.3850 .3675 .3113</td>
<td>.8475 .8063 .7712</td>
<td>.6175 .6375 .6125</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.3225 .3225 .2575</td>
<td>.8225 .7837 .7581</td>
<td>.6325 .6600 .6219</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.3175 .3250 .2581</td>
<td>.8200 .7588 .7244</td>
<td>.5475 .5713 .5869</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2450</td>
<td>.7087</td>
<td>.6525 .6012 .6094</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.2800</td>
<td>.1944</td>
<td>.7650 .7300 .6675</td>
<td>.5350 .5138 .5406</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2175</td>
<td>.1400</td>
<td>.7475 .7025 .6525</td>
<td>.5150 .5050 .5419</td>
</tr>
<tr>
<td>20/2 1000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.3200</td>
<td>.2425</td>
<td>.2188</td>
<td>.7500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2950</td>
<td>.2125</td>
<td>.2062</td>
<td>.7200</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.2550</td>
<td>.2125</td>
<td>.2037</td>
<td>.7400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2250</td>
<td>.1800</td>
<td>.1737</td>
<td>.6350</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.1850</td>
<td>.1450</td>
<td>.1750</td>
<td>.5950</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.1600</td>
<td>.1125</td>
<td>.1425</td>
<td>.5250</td>
</tr>
<tr>
<td>2,000</td>
<td>.9</td>
<td>$S-X^2$</td>
<td>.3950</td>
<td>.3525</td>
<td>.2950</td>
<td>.8350</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.3700</td>
<td>.3550</td>
<td>.2900</td>
<td>.8450</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.3600</td>
<td>.3000</td>
<td>.2600</td>
<td>.7700</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.3450</td>
<td>.2625</td>
<td>.2425</td>
<td>.7550</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-X^2$</td>
<td>.2800</td>
<td>.2375</td>
<td>.2200</td>
<td>.8050</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$S-G^2$</td>
<td>.2350</td>
<td>.2300</td>
<td>.2025</td>
<td>.7850</td>
</tr>
</tbody>
</table>

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.
Table 5. Summary of Comparison Among Different Multidimensional Structures.

<table>
<thead>
<tr>
<th>Test structure</th>
<th>HO-IRT</th>
<th>MIRT</th>
<th>Bifactor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image" alt="Diagram HO-IRT" /></td>
<td><img src="image" alt="Diagram MIRT" /></td>
<td><img src="image" alt="Diagram Bifactor" /></td>
</tr>
<tr>
<td>Formulation</td>
<td>$P(y_{jt} = 1</td>
<td>a_i, b_i, \theta_j) = \frac{1}{1 + \exp(-D(a_i \theta_j - b_i))}$</td>
<td>$P(y_{jt} = 1</td>
</tr>
<tr>
<td>Latent traits structure</td>
<td>$\xi \sim \text{MVN}[0, \begin{bmatrix} \lambda^2_1 &amp; \cdots &amp; \lambda^2_4 \ \cdots &amp; \cdots &amp; \cdots \ 0 &amp; \cdots &amp; \cdots \end{bmatrix}]$</td>
<td>$\theta_j \sim \text{MVN}[0, \begin{bmatrix} 0 \ \cdots \ 0 \end{bmatrix}]$</td>
<td>$\theta_0 \sim \text{MVN}[0, \begin{bmatrix} 0 \ \cdots \ 0 \end{bmatrix}]$</td>
</tr>
<tr>
<td>Notation</td>
<td>$a_i$, discrimination parameter</td>
<td>$a_i$, discrimination parameter</td>
<td>$a_D$, general discrimination parameter</td>
</tr>
<tr>
<td></td>
<td>$b_i$, difficulty parameter</td>
<td>$b_i$, difficulty parameter</td>
<td>$b_i$, discrimination parameter</td>
</tr>
<tr>
<td></td>
<td>$\theta_j$, domain-specific ability</td>
<td>$\theta_j$, ability</td>
<td>$\theta_j$, group ability</td>
</tr>
<tr>
<td></td>
<td>$\xi$, overall ability</td>
<td>$D$, a scaling constant</td>
<td>$D$, a scaling constant</td>
</tr>
<tr>
<td></td>
<td>$\epsilon_j$, residual</td>
<td>$(i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T)$</td>
<td>$(i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T)$</td>
</tr>
<tr>
<td></td>
<td>$D$, a scaling constant</td>
<td>$(i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T)$</td>
<td>$(i = 1, \ldots, I; j = 1, \ldots, N; t = 1, \ldots, T)$</td>
</tr>
<tr>
<td>#Parameter</td>
<td>$2 \times I + T$</td>
<td>$2 \times I + T \times (T - I) / 2$</td>
<td>$3 \times I$</td>
</tr>
</tbody>
</table>

Note. Take four-dimension structure as an example. HO-IRT = higher order item response theory; MIRT = multidimensional item response theory.
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 (\(\rho\)) 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), \(S-G^2\) had smaller FPRs than \(S-X^2\) under all conditions, which was consistent with Study 1, but the manipulated factors did not have a consistent effect on the FPRs. Regarding CDR, \(S-X^2\) had smaller CDRs than \(S-G^2\) 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 \(S-X^2\) and from .030 to .160 for \(S-G^2\). 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.

#### Table 6. Comparison Between \(S-X^2\) and \(S-G^2\) When the Misspecifying Items Have a Bifactor Structure.

<table>
<thead>
<tr>
<th>(\rho)</th>
<th>Index</th>
<th>10% FPR</th>
<th>10% CDR</th>
<th>20% FPR</th>
<th>20% CDR</th>
<th>40% FPR</th>
<th>40% CDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>(S-X^2)</td>
<td>.0847</td>
<td>.1100</td>
<td>.0894</td>
<td>.0825</td>
<td>.0950</td>
<td>.0869</td>
</tr>
<tr>
<td></td>
<td>(S-G^2)</td>
<td>.0397</td>
<td>.0525</td>
<td>.0453</td>
<td>.0362</td>
<td>.0454</td>
<td>.0400</td>
</tr>
<tr>
<td>.7</td>
<td>(S-X^2)</td>
<td>.0919</td>
<td>.0700</td>
<td>.1016</td>
<td>.1225</td>
<td>.1021</td>
<td>.1256</td>
</tr>
<tr>
<td></td>
<td>(S-G^2)</td>
<td>.0436</td>
<td>.0300</td>
<td>.0494</td>
<td>.0650</td>
<td>.0471</td>
<td>.0737</td>
</tr>
<tr>
<td>.5</td>
<td>(S-X^2)</td>
<td>.0942</td>
<td>.1400</td>
<td>.1166</td>
<td>.1375</td>
<td>.1225</td>
<td>.1431</td>
</tr>
<tr>
<td></td>
<td>(S-G^2)</td>
<td>.0400</td>
<td>.1600</td>
<td>.0537</td>
<td>.1388</td>
<td>.0579</td>
<td>.1275</td>
</tr>
</tbody>
</table>

Note. FPR = false positive rate; CDR = correct detection rate.
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 $S-X^2$ and $S-G^2$ 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 $S-X^2$ and $S-G^2$ 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, $S-G^2$ is recommended for HO-IRT models due to its smaller FPR and adequate CDR. Similar to the findings reported in the literature, both $S-X^2$ and $S-G^2$ perform poorly to detect model misspecification due to guessing behavior at the item level and perform well to detect different discrimination scales. Furthermore, both $S-X^2$ and $S-G^2$ perform too poorly to detect item misspecification due to multivariate structure of latent traits.

Because $S-X^2$ and $S-G^2$ 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
Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.


