Article

Modified Item-Fit Indices for Dichotomous IRT Models with Missing Data

Applied Psychological Measurement
2022, Vol. 46(8) 705–719

DOI: 10.1177/01466216221125176
journals.sagepub.com/home/apm
SAGE

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

Abstract

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 results will guide item revision and/or deletion (Liu & Maydeu-Olivares, 2014). During data collection, missing response data may likely happen due to various reasons. Chi-square-based item fit indices (e.g., Yen's Q_1, McKinley and Mill's G^2, Orlando and Thissen's $S\text{-}X^2$ and $S\text{-}G^2$) are the most widely used statistics to assess item-level fit. However, the role of total scores with complete data used in $S\text{-}X^2$ and $S\text{-}G^2$ is different from that with incomplete data. As a result, $S\text{-}X^2$ and $S\text{-}G^2$ cannot handle incomplete data directly. To this end, we propose several modified versions of $S\text{-}X^2$ and $S\text{-}G^2$ to evaluate item-level fit when response data are incomplete, named as $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$, of which the subscript "*impute*" denotes different imputation methods. Instead of using observed total scores for grouping, the new indices rely on imputed total scores by either a single imputation method or three multiple imputation methods (i.e., two-way with normally distributed errors, corrected item-mean substitution with normally distributed errors and response function imputation). The new indices are equivalent to $S\text{-}X^2$ and $S\text{-}G^2$ when response data are complete. Their performances are evaluated and compared via simulation studies; the manipulated factors include test length, sources of misfit, misfit proportion, and missing proportion. The results from simulation studies are consistent with those of Orlando and Thissen (2000, 2003), and different indices are recommended under different conditions.

Keywords

item fit, missing data, Chi-square-based item fit indices, multiple imputation, single imputation

Introduction

The existence of missing data has always been a challenge in psychometrics research. During data collection, missing response data may likely happen due to various reasons. Some of the resulting

[1]China Institue of Rural Education Development, Northeast Normal University, China
[2]College of Education, University of Washington, Washington, DC, USA

Corresponding Author:
Xue Zhang, China Institute of Rural Education Development, Northeast Normal University, 5268 Renmin Street, Changchun, Jilin Province, 130024, China.
Email: zhangx815@nenu.edu.cn

missing responses are partly due to special test design, under which some items are not administered and therefore missing by design (e.g., NEAP). Another type of missingness occur because of the respondent's behavior during the test, such as failing to reach the end of a test due to test speedness, or he/she decided to omit the item after reading it (Mislevy & Wu, 1996). From statistical perspective, the types of missing mechanism are summarized (Little & Rubin, 2002; Rubin, 1976) as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Under the MCAR assumption, the probability of missingness is the same for all respondents, such that the observed data are a simple random sample of the hypothetical complete data. Under the MAR assumption, the probability of missingness depends solely on the *observed* values of some other variables but not the missing value. Under the MNAR assumption, the probability of missing data on a particular variable depends on the missing variables such as possible missing responses or latent variables, even after controlling for the observed values of other variables (Enders & Baraldi, 2018). Mathematically, the missing mechanism can be expressed as follows,

$$p(\text{data} \mid \text{complete variables}) = p(\text{data} \mid \text{observed variables},\ \text{missing variables})$$

$$= \begin{cases} p(\text{data}), & \text{if data are MCAR,} \\ p(\text{data} \mid \text{observed variables}), & \text{if data are MAR,} \\ p(\text{data} \mid \text{observed variables},\ \text{missing variables}), & \text{if data are MNAR} \end{cases}$$

When the missing mechanism is MCAR or MAR, the missingness is deemed ignorable when to estimate parameters. On the contrary, the missingness is non-ignorable for MNAR and hence estimating model parameters solely on observed data would produce bias. To handle missing data that satisfy the MCAR or MAR assumptions, one popular way is listwise deletion, and another widely used way is to impute the missing data.

To evaluate item fit when response data are complete, numerous statistical procedures have been introduced in IRT literature. Among them, Chi-square-based item fit indices (e.g., Yen's Q_1 (1981), McKinley and Mill's G^2 (1985), Stone's χ^2* and G^2* (2000), Orlando and Thissen's $S\text{-}X^2$ and $S\text{-}G^2$) have been applied in various scenarios. For instance, they are used to examine model misspecification under dichotomous or/and polytomous items (Chon et al., 2010; Kang & Chen, 2008; Liang & Wells, 2009), to test violation of the monotonicity assumption of the item response function (IRF; Orlando & Thissen, 2003), to test item misfit due to Q-matrix misspecification (Wang et al., 2015), to identify item misfit in multidimensional or hierarchical item response models (Li & Rupp, 2011; Zhang & Stone, 2008; Zhang et al., 2018). Stone's χ^2* and G^2*, which can be treated as a Bayesian version of Q_1 and G^2, were used to detect item parameter drift (LaHuis et al., 2011; Stone & Zhang, 2003).

The posterior predictive model checking (PPMC) method (Sinharay, 2005; 2006) was also used to assess item fit as under a Bayesian estimation framework (e.g., Markov chain Monte Carlo). Furthermore, under the residual analysis framework, Haberman (2009) and Haberman et al. (2013) used generalized residuals to assess item fit. Above mentioned indices used different forms of discrepancy measures to quantify the discrepancy between model prediction and observation, on the other hand, the Lagrange multiplier (LM) test (Glas & Suárez-Falcón, 2003) was used to identify item misfit due to violation of local independence.

Moreover, the root integrated squared error (RISE, Douglas & Cohen, 2001) was presented to assess item fit for parametric IRT using nonparametric information, and the limited information fit statistics (Bartholomew & Leung, 2002; Cai et al., 2006; Maydeu-Olivares & Joe, 2005; Reiser, 2008) was proposed using the marginal tables to detect local dependence (Liu & Maydeu-Olivares, 2013) or identify the source of misfit (Liu & Maydeu-Olivares, 2014). Recently, in order to assess latent variable distribution fit in IRT, Li and Cai (2018) proposed Satorra–Bentler type moment adjustment of Pearson's χ^2, and compared with the unadjusted index and Maydeu-Olivares and Joe's M_2.

Previous studies examining the behavior of item fit indices have not mentioned incomplete response data. As a widely used family of item fit indices, current Chi-square-based item fit indices rely on person's θ estimates (e.g., Yen's Q_1 and McKinley and Mill's G^2) or total scores (e.g., Orlando and Thissen's $S\text{-}X^2$ and $S\text{-}G^2$) for grouping individuals. $S\text{-}X^2$ and $S\text{-}G^2$ outperform Q_1 and G^2 with completed data (e.g., Chon et al., 2010; Kang & Chen, 2008; Orlando & Thissen, 2000; 2003). As $S\text{-}X^2$ and $S\text{-}G^2$ rely on total scores, and the role of total scores for complete data is not the same as that for incomplete data, it may be problematic to assess item-level fit using Q_1, G^2, $S\text{-}X^2$ or $S\text{-}G^2$ for incomplete data. The primary goal of this study, thus, is to propose an effective item fit index to detect item misfit when the observed response matrix is incomplete.

This remainder of this article is organized as follows: first, we briefly review the IRT models. Then, the modified $S\text{-}X^2$ and $S\text{-}G^2$ indices, namely the $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ for incomplete dichotomous response data,[1] are proposed. Third, a simulation study is conducted to investigate the performances of the proposed indices to detect item misfit. Finally, we end with concluding remarks.

Method

Model Description

In this study, we consider a family of unidimensional binary IRT models (Baker & Kim, 2004): the one-parameter logistic (1PL), two-parameter logistic (2PL) and three-parameter logistic (3PL) models. The probability of examinee j ($j = 1, \ldots, N$) answers item i ($i = 1, \ldots, I$) correctly can be expressed as

$$p_{ij} = p\left(y_{ij} = 1 \middle| a_i, b_i, c_i, \theta_j\right) = c_i + \frac{1 - c_i}{1 + \exp\left[-D\left(a_i\theta_j - b_i\right)\right]}, \tag{1}$$

where y_{ij} denotes the response[2] of examinee j on item i, θ_j is the latent trait (i.e., ability) of examinee j, and a_i, b_i, c_i are the ith item's intercept, slope, and lower asymptote parameters, respectively. Here, the lower asymptote parameter can represent the guessing behavior. D is the scaling constant, which is set to 1 here. The formula of the 2PL model can be obtained by setting $c_i = 0$ for all i in equation (1), analogously, the formula of the 1PL model can be obtained by setting both $a_i = 1$ and $c_i = 0$ for all i in equation (1).

Accurate parameter estimation is the basis for statistical inference; biased parameter estimation may mislead the subsequent statistical analysis. When the missing data are missing not at random, the missing mechanism needs to be properly modeled during parameter estimation to avoid any potential bias. Although there exists a literature on modeling non-ignorable missing data with IRT (Debeer et al., 2017; Köhler et al., 2015; Liu & Wang, 2016; Lu & Wang, 2020; Rose et al., 2016), each modeling approach is only specific to one type of non-ignorable missing mechanism that it is hard to find a one-size-fits-all approach. As an exploratory research on item-level fit assessment with missing data, we will only focus on the simplest scenario (i.e., missing completely at random) in this study.

Latent Trait Estimation-Based Indices

Both Q_1 and G^2 are computed based on grouping individuals according to their latent trait estimation. Examinees are rank-ordered and partitioned into 10 homogeneous subgroups based on $\widehat{\theta}$. Then Q_1 and G^2 for item i can be written as

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

where k (k = 1, …, 10) represents a homogeneous group of examinees, N_k is the number of examinees in group k, O_{ik} and E_{ik} are the observed and expected proportions of correct responses for item i in group k. Here E_{ik} equals the mean predicted probability of a correct response from the model in each interval. The degrees of freedom (df) associated with Q_1 and G^2 are both $10 - m$, where m is the number of item parameters.

Number-Correct Score-Based Indices

Orlando and Thissen (2000) proposed item fit indices $S\text{-}X^2$ and $S\text{-}G^2$ based on number-correct scores (NC scores; i.e., total scores) for dichotomous items. The forms of $S\text{-}X^2$ and $S\text{-}G^2$ for item i are

$$S - X_i^2 = \sum_{k=1}^{I-1} N_k \frac{(O_{ik} - E_{ik})^2}{E_{ik}(1 - E_{ik})} \text{and } S - G_i^2 = 2\sum_{k=1}^{I-1} N_k \left[O_{ik} \ln \frac{O_{ik}}{E_{ik}} - (1 - O_{ik}) \ln \frac{1 - O_{ik}}{1 - E_{ik}} \right], \quad (3)$$

where k = 1, …, $I - 1$ denotes the subgroup, which is classified depend on NC scores, I denotes the test length. As the proportion of examinees who answered item i correctly is always 0 (or 1) at the two extremes (i.e., NC scores equals 0 or I), only $I - 1$ subgroups are considered to detect item misfit. Here N_k, O_{ik}, and E_{ik} have the same meanings as equation (2). The calculation of E_{ik} will be provided in the following subsection. The df associated with $S\text{-}X^2$ and $S\text{-}G^2$ are both $I - 1 - m$, where m is the number of item parameters for each item.

When response data are incomplete, classification of examinees based on NC scores may be mis-specified, as the role of NC scores for incomplete data is not the same as that for complete data. For instance, for a test with 20 items, one respondent answers 10 items correctly and 10 items incorrectly, his/her NC score is 10, another respondent answers 10 items correctly, 4 items incorrectly, and the responses of 6 items are missing, his/her NC score is also 10. Obviously, the ability of the second respondent may be higher than the first one with a high probability. Although their NC scores are the same, it may be problematic to classify them into the same subgroup.

Number-Correct Imputed Score-Based Indices

To handle incomplete data, we propose to first impute the missing data and then apply $S\text{-}X^2$ and $S\text{-}G^2$ to the imputed complete data. Specifically, the strategy for assessing item fit of an IRT model with/without missing data can be summarized as follows: (1) Estimate the item parameters based on a specified model; (2) impute the missing part of response data using different imputation methods (i.e., single imputation method, two-way with normally distributed errors (TWE), corrected item-mean substitution with normally distributed errors (CIMSE) and response function (RF) imputation),[3] then classify the examinees into K subgroups according to the "complete" observed total score; (3) calculate the observed and predicted frequencies of correctly/incorrectly responses for each item and each subgroup; (4) calculate $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ by computing the discrepancy between observed and predicted values.

The expressions of $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ for a dichotomous item i are as follows:

$$M_{impute} - X_i^2 = \sum_{k=1}^{I-1} \frac{(f_{ik} - N_k E_{ik})^2}{N_k E_{ik}(1 - E_{ik})}, \quad (4)$$

and

$$M_{impute} - G_i^2 = 2\sum_{k=1}^{I-1}\left[f_{ik}\ln\left(\frac{f_{ik}}{N_k E_{ik}}\right) - (N_k E_{ik} - f_{ik})\ln\left(\frac{N_k - f_{ik}}{N_k - N_k E_{ik}}\right)\right], \quad (5)$$

where N_k denotes the number of examinees in group k ($k = 1, \ldots, I$) using the imputed response data, and $I - 1$ subgroups are considered. f_{ik} denotes the number of examinees whose (imputed) response on item i is correct in the kth group, and E_{ik} denotes the expected proportion of correct response for item i in the kth group. Hence, $N_k E_{ik}$ is the corresponding predicted frequency of correctly responses. Please note that Equations (4) and (5) are essentially the same as Equation (3), except that the statistics are calculated using the imputed data, whereas $S\text{-}X^2$ and $S\text{-}G^2$ are computed based on the complete observed data. E_{ik} has the same form as $S\text{-}X^2$ and $S\text{-}G^2$,

$$E_{ik} = \frac{\int p_i(\theta) f^{*i}(k-1|\theta)\phi(\theta)d\theta}{\int f(k|\theta)\phi(\theta)d\theta}, \quad (6)$$

where $p_i(\theta)$ is the IRF of item i, $f(k|\theta)$ is the NC score likelihood given particular θ for score k, $f^{*i}(k-1|\theta)$ is the NC score likelihood given particular θ for score $k-1$ excluding item i, and $\phi(\theta)$ denotes the prior distribution of θ. The calculation is the same as that in Orlando and Thissen (2000).

The *df* associated with $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ both equal $I - 1 - m$ and their asymptotic distributions are the same, where m is the number of item parameters. Obviously, $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ are the general version of $S\text{-}X^2$ and $S\text{-}G^2$, which can handle both complete data and incomplete data.

Hereafter, Q_1, $S\text{-}X^2$, and $M_{impute}\text{-}X^2$ are named as χ^2–type indices as they rely on the χ^2 statistics; G^2, $S\text{-}G^2$, and $M_{impute}\text{-}G^2$ are named as *LR*-type indices as they rely on the likelihood ratio (LR) test.

Imputation methods

In this article, the single imputation method and three multiple imputation methods[4] (i.e., two-way with normally distributed errors, corrected item-mean substitution with normally distributed errors and response function imputation) are considered, as the single imputation method is the simplest method, and these three multiple imputation methods perform similarly and they are recommended by van Ginkel et al. (2007). These methods are introduced briefly below.

Single imputation method. As the simplest imputation method, the final imputed value of the missing response using the single imputation method is the row-mean score.

Two-way with normally distributed errors *(TW-E)*. Bernaards and Sijtsma (2000) defined $TW_{ij} = PM_j + IM_i - OM$ as the temporary value of the imputed score of examinee j on item i, where PM_j denotes the mean of the observed item scores of examinee j, IM_i is the mean of the observed item scores of item i, and OM denotes the mean of all observed item scores. A random error ε_{ij} was added to TW_{ij}, which was distributed normally with mean 0 and variance $\sum_{i,j\in obs}(y_{ij} - TW_{ij})^2/(\#obs - 1)$, where *obs* denotes the set of all observed cells (Bernaards & Sijtsma, 2000) and $\#obs$ is the corresponding size. Then let $TW_{ij}(E) = TW_{ij} + \varepsilon_{ij}$, the imputed value in cell (i, j) of the response matrix is obtained by rounding $TW_{ij}(E)$ to the nearest integer.

Corrected item-Mean Substitution with Normally Distributed Errors. The temporary imputed value in cell (i, j) is defined as (Huisman, 1998)

$$CIMS_{ij} = \left(\frac{PM_j}{\sum_{i \in obs(j)} IM_i \Big/ \#obs(j)} \right) \times IM_i,$$

where *obs(j)* denotes the set of observed cells for examinee j. A similar random error was added to $CIMS_{ij}$, which had the same distribution as that of $TW_{ij}(E)$. The final imputed value in cell (i, j) is obtained by rounding $CIMS_{ij}(E)$ to the nearest integer.

Response function imputation. Sijtsma and van der Ark (2003) defined $\widehat{R}_{(-i)j} = PM_j \times (I - 1)$ as the estimated value of the rest-score of examinee j on item i. Then, we can obtain $\mathrm{P}(y_{ij} = 1 | R_{(-i)j} = r) = \frac{\#(y_{i\cdot}=1, \widehat{R}_{(-i)\cdot}=r)}{\#(\widehat{R}_{(-i)\cdot}=r)}$ when r is an integer, and when r is not an integer, $\mathrm{P}(y_{ij} = 1 | R_{(-i)j} = r)$ can be approximated by the linear interpolation method. Interested readers can refer to Sijtsma and van der Ark (2003) for details. Then the imputed value in cell (i, j) can be drawn from a Bernoulli distribution with a successful probability $\mathrm{P}(y_{ij} = 1 | R_{(-i)j} = r)$.

Simulation Study

The main purpose of the simulation study is to examine the performances of M_{impute}-X^2 and M_{impute}-G^2, and to compare them with S-X^2, Q_1 and G^2. Hereafter, these four imputation methods are named as data-based imputation methods. Furthermore, treating all missing responses as wrong is also compared as an alternative imputation method.

For all conditions, the item parameters were calibrated via the expectation-maximization (EM) algorithm implemented by the R "mirt" package and the person parameters were obtained using the maximum likelihood estimation (MLE).

Simulation Design

This simulation study was designed to examine the performances of the proposed indices to differentiate mis-specified item response functions. Five factors and their varied conditions were considered: (1) test length (Orlando & Thissen, 2000),10 (small) and 40 (large); (2) sample size (Orlando & Thissen, 2000), 500 (small) and 1000 (large); (3) missing proportion, 0.01 (extra small), 0.1 (medium), and 0.2 (large); (4) item misfit proportion (Wang et al., 2015), 0.1 (small) and 0.4 (large); and (5) five different true data generation models and three different calibration models (details were provided in Table 1). The missing mechanism considered in this study was missing at random completely. To generate the response data, the distributions used to generate item parameters were presented in the 4*th* column in Table 1, and the latent traits were simulated from standard normal $N(0, 1)$. 500 replications were done for each condition.

The first three generation models are a mixture of 1PL and 2PL models, a mixture of 1PL and 3PL models, and a mixture of 2PL and 3PL models, in which the misfit items conform to the more complex model, and the response data are fitted using the other one. In other words, well-fit items conform to the simpler model.

The fourth type of misfit items conforms to a 4-parameter logistic (4PL) model (Barton & Lord, 1981), in other words, the probability of answering correctly will never reach 1. It is expressed by

$$p(y = 1 | a, b, c, d, \theta) = c + \frac{d - c}{1 + \exp[-D(a\theta - b)]}, \tag{7}$$

Table 1. Summary of simulation design.

Scenario	Well fit item	Misfit item	Parameter generation
1	$p = \frac{1}{1+\exp[-D(\theta-b)]}$	$p = \frac{1}{1+\exp[-D(a\theta-b)]}$	$a \sim \log N(0, 0.5)$ $b \sim \log N(0, 0.5) \times N(0, 1)$ $c \sim \text{logit}N(-1.1, 0.5)$ $d \sim Beta(8, 2)$ $e \sim \log N(0, 0.5) \times U(0, 2)$
2	$p = \frac{1}{1+\exp[-D(\theta-b)]}$	$p = c + \frac{1-c}{1+\exp[-D(a\theta-b)]}$	
3	$p = \frac{1}{1+\exp[-D(a\theta-b)]}$	$p = c + \frac{1-c}{1+\exp[-D(a\theta-b)]}$	
4	$p = c + \frac{1-c}{1+\exp[-D(a\theta-b)]}$	$p = c + \frac{d-c}{1+\exp[-D(a\theta-b)]}$	
5	$p = c + \frac{1-c}{1+\exp[-D(a\theta-b)]}$	$p = \frac{c}{1+\exp\{D[a\theta-(b-e)]\}} + \frac{1}{1+\exp\{D(a\theta-b)\}}$	

where a, b, c, and D are the same as in Equation (1), d denotes the upper asymptote parameter, which is drawn from $Beta(8, 2)$. The upper asymptote parameter can represent the slipping behavior. The response data are fitted using a 3PL model.

The fifth type of misfit items violates the monotonicity assumption of IRF, the probability of a correct response is described by Thissen (1986) and Wainer and Thissen (1987) and can be expressed by

$$p(y = 1|a, b, c, d, \theta) = \frac{c}{1+\exp\{D[a\theta-(b-e)]\}} + \frac{1}{1+\exp\{D(a\theta-b)\}}, \tag{8}$$

where a, b, c, and D are the same as in Equation (1), e is a positive number, which is drawn from $\log N(0, 0.5)\times U(0, 2)$, and larger e leads to a larger dip in the curve. Here, the response data are also fitted using a 3PL model.

To evaluate and compare the performances of the item fit indices, false positive rates (FPRs) and correct detection rates (CDRs) are calculated with $\alpha = 0.05$. The FPR is defined as the proportion of well-fit items that are mistakenly flagged, and it is computed per replication and averaged over 500 replications. The CDR is defined as the proportion of misfit items that are correctly detected, and it is also computed per replication and averaged over 500 replications. Moreover, 95% confidence intervals (CIs) for the true rejection rates are reported to account for sampling error associated with expected rejecting rates

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

where R is the number of replications and α is the significant level (which is set to .05). In this study, the expected 95% CI is [0.031, 0.069].

Simulation Results

Comparison of different imputation methods. To further evaluate the performances of different imputation methods, Analysis of Variance (ANOVA)[5] is done for the p-values of modified item fit indices when test length is small, as the 500 replications can be treated as 500 independent observations. Furthermore, when the null hypothesis is rejected, post hoc multiple comparison test is done to find out which methods differ. The effect size measures (i.e., η^2) for small test length cases are summarized in the online supplement, and the average η^2 is calculated respectively for

misfit and well fit items and is presented in Table 2. Hereafter, *P.mix* denotes the mixed proportion, *P.miss* denotes the missing proportion.

When the missing proportion was extra small, η^2 is equal to or approximately equals to 0 expect for misfit items in Scenarios 4–5. For other conditions, η^2s for misfit items are smaller than those for well fit items in Scenarios 1–2, but larger in Scenarios 3–5. Larger missing proportion leads to larger η^2.

In addition, based on the results from ANOVA and post hoc multiple comparisons, it can be concluded that when the missing proportion is extra small, there are no significant differences among four data-based imputation methods for misfit items in Scenarios 4–5, and there are no significant differences among all the imputation methods for other scenarios and well fit items in Scenarios 4–5. The difference between the performance of treating all missing responses as wrong and performances of data-based imputation methods are significant.

When the missing proportion is medium, there is no significant difference among four data-based imputation methods for misfit items in Scenarios 1–3 and well fit items in Scenarios 4–5; there is no significant difference between TW-E and RF or among SI, CIMS-E and RF for well fit items in Scenario 1; the difference between CIMS-E imputation method and other data-based imputation methods is significantly for well fit items in Scenarios 2–3; the difference between SI method and other data-based imputation methods is significant for misfit items in Scenarios 4–5. When the missing proportion is large, CIMS-E is significantly different from other data-based imputation methods for misfit items in Scenario 1 and well fit items in Scenarios 1–3; there is significant difference between SI and other imputation methods for misfit items in Scenarios 3–5; there is no significant difference among all the data-based imputation methods for most misfit items in Scenarios 2 and well fit items in Scenarios 4–5.

Overall, the differences between TW-E and RF are not significant for the bulk of items in the small test length cases. Obviously, the type of imputation method has little/no effect when the missing proportion is extra small, as under this condition there are few differences among different types of NC imputed scores.

Table 2. Summary of average η^2 for M_{impute}-X^2 with small test length.

			Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
N	*P. mix*	*P. miss*	*Misfit*	*Fit*	*Misfit*	*Fit*	*Misfit*	*Fit*	*Misfit*	*Fit*	*Misfit*	*Fit*
500	0.1	0.01	0	.000	.002	.000	.002	.000	.048	.001	.030	.001
		0.1	.010	.015	.048	.022	.115	.033	.148	.046	.105	.035
		0.2	.035	.054	.062	.065	.189	.057	.190	.065	.163	.063
	0.4	0.01	.000	.000	.001	.001	.002	.001	.031	.001	.038	.001
		0.1	.008	.014	.038	.045	.070	.071	.099	.095	.107	.061
		0.2	.030	.047	.058	.123	.111	.092	.121	.110	.155	.086
1000	0.1	0.01	0	.000	.002	.000	.003	.001	.081	.001	.040	.002
		0.1	.009	.032	.048	.043	.159	.054	.114	.077	.105	.071
		0.2	.042	.078	.056	.103	.215	.070	.206	.079	.139	.137
	0.4	0.01	0	.000	.001	.002	.002	.001	.045	.002	.046	.002
		0.1	.013	.031	.048	.085	.104	.110	.101	.145	.108	.098
		0.2	.036	.073	.058	.152	.136	.095	.105	.120	.114	.092

Note. "*Misfit*" denotes the average of η^2 for misfit items, and "*Fit*" denotes the average of η^2 for well-fit items.

False positive rate and correct detection rate

Tables 3 and 4 display the comparison among different χ^2–based indices (i.e., Q_1, $S\text{-}X^2$ and five $M_{impute}\text{-}X^2$) through FPRs and CDRs, respectively. In these tables, the entries under the $M\text{-}X^2$ column are the smallest FRPs (in Table 3) among different $M_{impute}\text{-}X^2$ and the corresponding CDRs (in Table 4). The full results of FPRs and CDRs are shown in the online supplementary. Obviously, $S\text{-}X^2$ cannot handle the cases with large test length and medium to large missing proportion. The highest number of entries of FPRs contained in 95% CI was in Scenario 2, and the lowest number was in Scenarios 4.

When the response data are generated by the 1PL model mixed with the 2PL model (i.e. Scenario 1), the values presented in Table 3 were mostly from $M_{SI}\text{-}X^2$ or $M_{TWE}\text{-}X^2$. Larger missing proportion leads to larger FPRs of $S\text{-}X^2$ and $M\text{-}X^2$ with small test length and smaller FPR with large test length, the corresponding FPR of Q_1 reverses. In terms of CDR, larger missing proportion leads to smaller CDRs of all indices except for $S\text{-}X^2$ with not small missing proportions and large test length. Q_1 outperformed when test length was large, but underperformed when test length is small.

When the response data is generated by the 1PL model mixed with the 3PL model (i.e., Scenario 2), the trends of FPRs and CDRs are similar to those in Scenario 1, but there are more entries of FPRs which are contained in 95% CI. The FPRs in the $M\text{-}X^2$ column are mostly from $M_{SI}\text{-}X^2$ or $M_{TWE}\text{-}X^2$. The modified indices outperform compared to other indices, the FPRs of Q_1 with small test length and extra small missing proportion are too large. Regarding with CDRs, $S\text{-}X^2$ underperforms compared to modified indices.

When the generation model is the 2PL model mixed with the 3PL model (i.e., Scenario 3), the values in the $M\text{-}X^2$ column are mostly from $M_{SI}\text{-}X^2$, $M_{TWE}\text{-}X^2$, or $M_{RF}\text{-}X^2$. The trends of FPRs in missing proportion are almost similar to Scenario 1, but larger missing proportion leads to larger CDRs of all types of indices with small test length and smaller CDRs of indices except $S\text{-}X^2$ with large test length. The FPRs of Q_1 are extremely large for small test length and mostly smallest for large test length. And the suitable values of CDRs are too small, which means that it is hard to distinguish the difference between the 2PL model and the 3PL model.

In Scenarios 4–5, most of values in the $M\text{-}X^2$ column are from $M_{SI}\text{-}X^2$, especially when test length is large. The trends are both similar to Scenario 3. Comparing to Scenario 3, the FPRs and CDRs become larger for small test length and smaller for large test length. Note that the performance of SI method is significantly different from multiple imputation methods for misfit items, though the results of $M_{SI}\text{-}X^2$ were presented in Tables 3 and 4, this index is not recommended due to its extreme small CDR.

In sum, Q_1 performs respectably when test length is large. Comparing the performances of item fit indices among Scenarios 1–3, they perform the best to distinguish 1PL model and 3PL model, and worst to distinguish 2PL model and 3PL model. When the scenarios become more complex (i.e., Scenarios 4–5), their performances to distinguish two different models becomes better.

Comparison of χ^2–type indices and LR-type indices

In Scenario 1, the FPR of Q_1 is smaller than G^2 when test length is small, and similar or larger than G^2 when test length is large, the CDRs of Q_1 and G^2 are similar. In Scenario 2, Q_1 has smaller FPR and larger CDR than G^2. In Scenario 3, the FPR of Q_1 is similar to or larger than G^2 with small test length, and smaller than G^2 with large test length, their CDRs are similar. And in Scenarios 4–5,

Table 3. Summary of FPRs using X^2–type indices.

			Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
J	*P. mix*	*P. miss*	Q_1	*S*-X^2	*M*-X^2	Q_1	*S*-X^2	*M*-X^2	Q_1	*S*-X^2	*M*-X^2	Q_1	*S*-X^2	*M*-X^2	Q_1	*S*-X^2	*M*-X^2
N = 500																	
10	0.1	0.01	.312	**.052**	**.059**	.352	**.058**	**.051**	.416	**.057**	**.058**	.824	.078	.08	.816	.084	.083
		0.1	.138	**.067**	.091	.141	.07	**.066**	.427	.096	.115	.832	.172	.129	.802	.156	.123
		0.2	.125	.083	.107	.131	.088	.091	.472	.151	.176	.852	.303	.203	.818	.29	.193
	0.4	0.01	.329	**.056**	.081	.481	**.053**	**.056**	.602	.05	**.058**	.945	.105	.092	.861	.081	.079
		0.1	.134	**.066**	.083	.126	**.064**	**.062**	.565	.099	.123	.917	.169	.145	.835	.149	.112
		0.2	.119	.091	.12	.132	.078	.077	.567	.16	.2	.921	.325	.192	.835	.292	.183
40	0.1	0.01	**.052**	**.061**	.096	**.046**	**.058**	.07	**.058**	**.065**	.083	.118	.095	.075	.119	.088	.077
		0.1	**.054**	.762	.091	**.046**	.745	**.067**	**.062**	.977	.075	.144	.998	.07	.132	.998	**.068**
		0.2	**.058**	I	**.068**	**.053**	I	**.048**	**.065**	I	**.058**	.17	I	.059	.168	I	**.057**
	0.4	0.01	**.057**	**.066**	.098	.078	.051	**.047**	**.068**	.074	.085	.151	.091	.071	.144	.087	.078
		0.1	**.06**	.77	.088	.072	.731	**.041**	.071	.98	.081	.182	.998	.071	.164	.998	.074
		0.2	**.062**	I	.071	.078	I	**.034**	.08	I	**.067**	.215	I	**.065**	.19	I	**.057**
N = 1000																	
10	0.1	0.01	.57	**.047**	**.058**	.625	**.055**	**.053**	.626	**.057**	**.051**	.934	.089	.08	.929	**.068**	.075
		0.1	.31	.077	.085	.36	.075	**.069**	.706	.085	.126	.93	.145	.129	.916	.14	.122
		0.2	.236	.075	.128	.275	.077	.115	.71	.115	.241	.928	.215	.244	.925	.215	.244
	0.4	0.01	.614	**.065**	.087	.786	**.058**	.071	.795	**.055**	.07	.984	.088	.077	.954	**.066**	.077
		0.1	.32	**.066**	.103	.305	.072	**.064**	.76	.098	.118	.974	.157	.134	.933	.13	.127
		0.2	.231	.079	.156	.297	.079	.102	.779	.131	.253	.973	.231	.269	.927	.208	.269
40	0.1	0.01	**.055**	**.063**	.096	**.048**	**.059**	**.067**	.072	**.058**	.078	.154	.07	.073	.155	.082	.07
		0.1	**.063**	.452	.093	**.051**	.447	**.063**	.076	.809	.082	.18	.92	.069	.18	.952	.07
		0.2	.076	I	.087	**.059**	I	**.059**	.083	I	.072	.239	I	**.065**	.239	I	**.064**
	0.4	0.01	**.064**	**.057**	.091	.163	**.05**	**.053**	.1	**.068**	.084	.207	.09	.071	.193	.078	.077
		0.1	**.067**	.467	.096	.149	.425	**.053**	.103	.808	.088	.278	.942	.073	.278	.949	.076
		0.2	.077	I	.088	.15	I	**.044**	.128	I	.085	.373	I	.07	.341	I	**.066**

Note. Boldface entries are the FPRs which are contained in 95% CI. The entries under the *M*-X^2 column are the smallest values among FPRs of different M_{impute}-X^2.

Table 4. Summary of CDRs using X^2–type indices.

			Scenario 1			Scenario 2			Scenario 3			Scenario 4			Scenario 5		
J	*P. mix*	*P. miss*	Q_1	S-X^2	M-X^2	Q_1	S-X^2	M-X^2	Q_1	S-X^2	M-X^2	Q_1	S-X^2	M-X^2	Q_1	S-X^2	M-X^2
N = 500																	
10	0.1	0.01	.536	.31	.326	.706	.408	.578	.282	.07	.044	.514	.096	.08	.56	.106	.098
		0.1	.312	.198	.245	.564	.248	.534	.494	.11	.082	.536	.136	.348	.57	.208	.402
		0.2	.242	.126	.238	.494	.17	.478	.576	.146	.108	.462	.298	.17	.488	.426	.264
	0.4	0.01	.525	.332	.288	.56	.401	.344	.343	.081	.057	.545	.1	.115	.584	.08	.093
		0.1	.296	.18	.284	.262	.241	.328	.343	.117	.099	.5	.15	.143	.593	.2	.389
		0.2	.249	.116	.262	.339	.174	.309	.406	.157	.104	.487	.281	.173	.55	.406	.271
40	0.1	0.01	.422	.243	.303	.575	.407	.554	.099	.074	.064	.162	.09	.024	.198	.113	.02
		0.1	.399	.768	.245	.558	.752	.421	.095	.977	.044	.163	.997	.029	.199	1	.037
		0.2	.351	—	.158	.515	—	.272	.074	—	.036	.219	—	.025	.211	—	.025
	0.4	0.01	.423	.265	.299	.425	.331	.384	.078	.09	.059	.175	.099	.081	.13	.118	.023
		0.1	.39	.762	.228	.383	.751	.295	.071	.974	.05	.18	.999	.035	.139	.999	.028
		0.2	.357	—	.143	.355	—	.194	.07	—	.037	.197	—	.042	.146	—	.029
N = 1000																	
10	0.1	0.01	.722	.476	.42	.888	.538	.652	.52	.088	.068	.67	.112	.074	.724	.076	.092
		0.1	.548	.314	.42	.794	.362	.662	.762	.08	.17	.646	.146	.362	.722	.138	.37
		0.2	.456	.132	.386	.762	.18	.716	.84	.11	.19	.672	.204	.692	.712	.282	.764
	0.4	0.01	.774	.473	.442	.821	.536	.499	.568	.093	.063	.684	.086	.123	.743	.079	.112
		0.1	.562	.287	.425	.503	.357	.579	.578	.111	.133	.657	.156	.482	.702	.171	.507
		0.2	.441	.156	.394	.633	.211	.485	.659	.144	.23	.629	.208	.618	.697	.288	.41
40	0.1	0.01	.557	.396	.415	.735	.539	.659	.15	.101	.082	.224	.083	.021	.296	.109	.017
		0.1	.511	.47	.331	.72	.512	.55	.148	.801	.072	.247	.942	.029	.305	.98	.024
		0.2	.503	—	.285	.672	—	.43	.139	—	.056	.287	—	.031	.343	—	.038
	0.4	0.01	.576	.393	.437	.586	.469	.509	.112	.103	.066	.236	.099	.121	.206	.098	.027
		0.1	.538	.482	.363	.557	.476	.414	.099	.79	.059	.25	.926	.037	.214	.976	.028
		0.2	.52	—	.284	.533	—	.309	.102	—	.046	.267	—	.04	.228	—	.03

Note. -denotes null. The entries under the M-X^2 column are the corresponding CDRs to the FPRs in Table 3, which are the smallest values among FPRs of different M_{impute}-X^2.

both FPR and CDR of Q_1 are larger than G^2. The modified χ^2–type indices mostly have larger FPRs and CDRs than the modified LR^2–type indices, the differences between their CDRs are quite apparent with small test length in Scenarios 2 and 5. It appears that most of LR^2–type indices have smaller FPRs and smaller CDRs than χ^2–type indices.

Conclusions

Data collection cannot in the nature of things be free from missingness for various reasons. Current Chi-square-based item fit indices rely on either latent trait estimation or total score. However, when the response data are incomplete, it is unreasonable to compare examinees' total scores without any pretreatment. To this end, we modify $S\text{-}X^2$ and $S\text{-}G^2$, which perform well to assess item-level fit for complete data (Orlando & Thissen, 2000; 2003; Wang et al., 2015; Zhang et al., 2018), to fill in the gap using different data imputation methods.

Across all simulation conditions, the modified indices using two-way with normally distributed errors imputation are recommended for UIRT models due to their appropriate false positive rates, acceptable correct detection rates and insignificant difference with other data-based imputation methods. When test length is large, the performances of Q_1 and G^2 are also recommended because of easy operating and acceptable FPRs and CDRs. When missing proportion is extra small, any item fit indices mentioned in this study can be used. Compared to $S\text{-}X^2$, all $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ with data-based imputation methods perform more stably in terms of FPRs. $M_{impute}\text{-}X^2$ and $M_{impute}\text{-}G^2$ with multiple imputation methods perform well to detect model misspecification due to non-invariant discrimination parameters, non-monotonic ICCs and slipping behavior, and perform poorly to detect model misspecification due to guessing behavior. The performances of $M_{SI}\text{-}X^2$ and $M_{SI}\text{-}G^2$ are not significantly different from the modified indices with multiple imputation method to detect model misspecification due to guessing/slipping behavior.

The missing mechanism considered in this article is missing completely at random, we use the NC imputed score as the matching criterion when constructing the discrepancy measure, which can be easily extended to deal with a variety of missing data scenarios. Furthermore, as the utilization of $S\text{-}X^2$ in many areas were reported in the literature, it is easy to extend this study to deal with polytomous data (Kang & Chen, 2008), multiple choice items (Thissen & Steinberg, 1984) and different multivariate factor structures (Zhang & Stone, 2008; Li & Rupp, 2011; Zhang et al., 2018). In addition, it would be worthwhile to examine the performances of the proposed indices to detect other sources of misfit. Finally, the performance of other item fit statistics (e.g., residual analysis, LM test, and PPMC) in the presence of missing data should be investigated in future research.

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: Humanities and Social Sciences Youth Foundation (Ministry of Education of the People's Republic of China): 20YJC880124, and IES R305D200015.

ORCID iDs

Xue Zhang https://orcid.org/0000-0002-3900-0118
Chun Wang https://orcid.org/0000-0003-2695-9781

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The subscript "*impute*" represents different imputation methods, which will be introduced in later sections.
2. In this study, the intercept/slope parameterization is used, which is equivalent to the discrimination/difficulty parameterization.
3. The detailed imputation methods were described in next subsection.
4. Notations are deferred to Equation (1).
5. Null hypothesis (H_0) is the performances of these imputation methods are equivalent.

References

Baker, F. B., & Kim, S. H. (Eds), (2004). *Item response theory: Parameter estimation techniques*. CRC Press

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), 1–15. https://doi.org/10.1348/000711002159617

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *ETS Research Report Series*, *1*(1), i–8. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Bernaards, C. A., & Sijtsma, K. (2000). Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. *Multivariate Behavioral Research*, *35*(3), 321–364. https://doi.org/10.1207/S15327906MBR3503_03

Cai, L., Maydeu Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited information goodness of fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173–194. https://doi.org/10.1348/000711005X66419

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*(3), 318–338. https://doi.org/10.1111/j.1745-3984.2010.00116.x

Debeer, D., Janssen, R., & De Boeck, P. (2017). Modeling Skipped and not reached items using IRTrees. *Journal of Educational Measurement*, *54*(3), 333–363. http://doi.org/10.1111/jedm.12147

Douglas, J., & Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, *25*(3), 234–243. https://doi.org/10.1177/01466210122032046

Enders, C. K., & Baraldi, A. N. (2018). Missing data handling methods. In: *The Wiley handbook of psychometric testing. A multidisciplinary reference on survey, scale and test development* (pp. 139–185).

Glas, C. A., & Suárez-Falcón, J. C. (2003). A comparison of item-fit statistics for the three- parameter logistic model. *Applied Psychological Measurement*, *27*(2), 87–106. https://doi.org/10.1177/0146621602250530

Haberman, S. J. (2009). Use of generalized residuals to examine goodness of fit of item response models. *ETS Research Report Series*, *1*(1), i-17. https://doi.org/10.1007/s11336-012-9305-1

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*(3), 417–440. https://doi.org/10.1007/s11336-012-9305-1

Huisman, M. (1998). *Item nonresponse: Occurrence, causes, and imputation of missing answers to test items*. Leiden, The Netherlands: DSWO Press.

Kang, T., & Chen, T. T. (2008). Performance of the generalized S X2 item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Köhler, C., Pohl, S., & Carstensen, C. H. (2015). Taking the missing propensity into accountwhen estimating competence scores: Evaluation of item response theory models for nonignorable omissions. *Educational and Psychological Measurement*, *75*(5), 850–874. http://doi.org/10.1177/0013164414561785

LaHuis, D. M., Clark, P., & O'Brien, E. (2011). An examination of item response theory itemfit indices for the graded response model. *Organizational Research Methods*, *14*(1), 10–23. https://doi.org/10.1177/1094428109350930

Li, Y., & Rupp, A. A. (2011). Performance of the S− χ2 statistic for full-information bifactor models. *Educational and Psychological Measurement*, *71*(6), 986–1005. https://doi.org/10.1177/0013164410392031

Li, Z., & Cai, L. (2018). Summed score likelihood–based indices for testing latent variable distribution fit in item response theory. *Educational and Psychological Measurement*, *78*(5), 857–886. https://doi.org/10.1177/0013164417717024

Liang, T., & Wells, C. S. (2009). A model fit statistic for generalized partial credit model. *Educational and Psychological Measurement*, *69*(6), 913–928. https://doi.org/10.1177/0013164409332222

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley.

Liu, C. W., & Wang, W.-C. (2016). Unfolding IRT models for likert-type items with a don't know option. *Applied Psychological Measurement*. *40*(7), 1–17. http://doi.org/10.1177/0146621616664047

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. *Educational and Psychological Measurement*, *73*(2), 254–274. https://doi.org/10.1177/0013164412453841

Liu, Y., & Maydeu-Olivares, A. (2014). Identifying the source of misfit in item response theory models. *Multivariate Behavioral Research*, *49*(4), 354–371. https://doi.org/10.1080/00273171.2014.910744

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement*, *57*(1), 584–620. https://doi.org/10.1111/jedm.12270

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*(471), 1009–1020. https://doi.org/10.1198/016214504000002069

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*(1), 49–57. https://doi.org/10.1177/014662168500900105

Mislevy, R. J., & Wu, P. K. (1996). Missing responses and IRT ability estimation: Omits, choice, time limits, and adaptive testing. *ETS Research Report Series*, *1996*(2), i–36. https://doi.org/10.1002/j.2333-8504.1996.tb01708.x

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, *24*(1), 50–64. https://doi.org/10.1177/01466216000241003

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*(4), 289–298. https://doi.org/10.1177/0146621603027004004

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology*, *61*(2), 331–360. https://doi.org/10.1348/000711007X204215

Rose, N., Davier, M., & Nagengast, B. (2016). Modeling omitted and not-reached items in IRT models. *Psychometrika*, *82*(3), 795–819. http://doi.org/10.1007/s11336-016-9544-7

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Sijtsma, K., & Van der Ark, L. A. (2003). Investigation and treatment of missing item scores in test and questionnaire data. *Multivariate Behavioral Research*, *38*(4), 505–528. https://doi.org/10.1207/s15327906mbr3804_4

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*(4), 375–394. https://doi.org/10.1111/j.1745-3984.2005.00021.x

Sinharay, S. (2006). Bayesian item fit analysis for unidimensional item response theory models. *British journal of mathematical and statistical psychology*, *59*(2), 429–449. https://doi.org/10.1348/000711005X66888

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*(1), 58–75. https://doi.org/10.1111/j.1745-3984.2000.tb01076.x

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*(4), 331–352. https://doi.org/10.1111/j.1745-3984.2003.tb01150.x

Thissen, D. (1986). Non-monotonic item characteristic curves. Invited presentation at the annual meeting of the American Education Association, San Francisco

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501–519. https://doi.org/10.1007/BF02302588

van Ginkel, J. R., van der Ark, L. A., & Sijtsma, K. (2007). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research*, *42*(2), 387–414. https://doi.org/10.1080/00273170701360803

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal ofEducational Statistics*, *12*(4), 339–368. https://doi.org/10.3102/10769986012004339

Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA. *Applied Psychological Measurement*, *39*(7), 525–538. https://doi.org/10.1177/0146621615583050

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*(2), 245–262. https://doi.org/10.1177/014662168100500212

Zhang, B., & Stone, C. A. (2008). Evaluating item fit for multidimensional item response models. *Educational and Psychological Measurement*, *68*(2), 181–196. https://doi.org/10.1177/0013164407301547

Zhang, X., Wang, C., & Tao, J. (2018). Assessing item-level fit for higher order item response theory models. *Applied Psychological Measurement*, *42*(8), 644–659. https://doi.org/10.1177/0146621618762740