

Evaluating the Fit of Sequential G-DINA Model Using Limited-
Information Measures

Wenchao Ma

Deposited 2023-09-27

Citation of published version:

Ma, W. (2019). Evaluating the Fit of Sequential G-DINA Model Using Limited-
Information Measures. In *Applied Psychological Measurement* (Vol. 44, Issue 3,
pp. 167–181). SAGE Publications. <https://doi.org/10.1177/0146621619843829>

Evaluating the Fit of Sequential G-DINA Model Using Limited-Information Measures

Applied Psychological Measurement
2020, Vol. 44(3) 167–181
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621619843829
journals.sagepub.com/home/apm



Wenchao Ma¹ 

Abstract

Limited-information fit measures appear to be promising in assessing the goodness-of-fit of dichotomous response cognitive diagnosis models (CDMs), but their performance has not been examined for polytomous response CDMs. This study investigates the performance of the M_{ord} statistic and standardized root mean square residual (SRMSR) for an ordinal response CDM—the sequential generalized deterministic inputs, noisy “and” gate model. Simulation studies showed that the M_{ord} statistic had well-calibrated Type I error rates, but the correct detection rates were influenced by various factors such as item quality, sample size, and the number of response categories. In addition, the SRMSR was also influenced by many factors and the common practice of comparing the SRMSR against a prespecified cut-off (e.g., .05) may not be appropriate. A set of real data was analyzed as well to illustrate the use of M_{ord} statistic and SRMSR in practice.

Keywords

cognitive diagnosis, ordinal response, model-data fit, goodness-of-fit, limited information, sequential G-DINA

Introduction

Cognitive diagnosis models (CDMs) aim to classify individuals into homogeneous latent classes. Each latent class has a unique profile of attributes, which are typically binary latent variables, representing the presence or absence of latent constructs of interest. A large number of CDMs have been developed in the literature. Examples of CDMs for dichotomous response include the deterministic inputs, noisy “and” gate (DINA; Haertel, 1989) model, the deterministic input noisy “or” gate (DINO; Templin & Henson, 2006) model, the *additive* CDM (A-CDM; de la Torre, 2011), and the generalized DINA (G-DINA; de la Torre, 2011) model. Examples of CDMs for polytomous responses include the sequential G-DINA model (sG-DINA; Ma & de la Torre, 2016), the diagnostic model for ordinal response (R. Liu & Jiang, 2018) and the general diagnostic model (von Davier, 2008). The usefulness of these CDMs,

¹The University of Alabama, Tuscaloosa, AL, USA

Corresponding Author:

Wenchao Ma, The University of Alabama, Box 870231, Tuscaloosa, AL 35487, USA.
Email: wenchao.ma@ua.edu

however, depends on whether they can adequately fit the data, and therefore, empirically examining the model-data fit is critical.

Traditional Pearson χ^2 statistic and likelihood ratio statistic G^2 are well recognized to be less useful in practice in that they consider the expected and observed frequencies of all response patterns. To address this issue, some limited-information measures based on the observed and predicted marginal frequencies of response patterns have been proposed (e.g., Reiser, 2008; Maydeu-Olivares & Joe, 2006). In CDMs, the M_2 statistic for dichotomous response has been shown to have well-calibrated Type I error rates under varied conditions and adequate power in detecting some types of model misspecifications (F. Chen, Liu, Xin, & Cui, 2018; Hansen, Cai, Monroe, & Li, 2016; Jurich, 2014; Y. Liu, Xin, Li, Tian, & Liu, 2016). The M_2 statistic can also be applied to graded response data, but its calculation may be difficult or even impossible due to a heavy computation burden when the number of items and response categories increase (Maydeu-Olivares & Joe, 2014). Building upon the M_2 statistic, Maydeu-Olivares (2013) and Cai and Hansen (2013) introduced the M_{ord} statistic (referred to as the M_2^*) for polytomous response item response models. They also found that the M_{ord} statistic had better-calibrated Type I error and higher power than the M_2 statistic in detecting model misspecifications for graded response data, especially when the number of categories was large. However, the performance of M_{ord} statistic for ordinal response CDMs has not been investigated.

In addition to the limited-information statistics with known limiting distributions, several limited-information indices have also been proposed as effect size measures. The root mean square error approximation (RMSEA) based on the M_2 statistic using the univariate and bivariate margins, typically referred to as RMSEA_2 , has been examined in several studies (e.g., Maydeu-Olivares & Joe, 2006; Y. Liu, Xin, et al., 2016). However, RMSEA_2 is a function of the number of categories and may not be suitable for polytomous response models (Maydeu-Olivares & Joe, 2014). Maydeu-Olivares and Joe (2014) recommended the use of the standardized root mean squared residual (SRMSR), and suggested that a model with $\text{SRMSR} < 0.05$ can be viewed as a well-fitted model. This criteria has been used in several studies on CDMs (Jiang & Ma, 2018; R. Liu, Huggins-Manley, & Bulut, 2018), but its appropriateness has not been examined.

This study aims to investigate the performance of the M_{ord} statistic and SRMSR for the sG-DINA model, which provides a general model framework to handle polytomously scored items that can be decomposed into a set of tasks and are scored sequentially. In addition, unlike other CDMs for polytomous responses, the sG-DINA model can account for the fact that different attributes may be involved in different tasks, and thus has the potential to provide more accurate estimation of students' attribute profiles.

Overview of the Sequential G-DINA Model

Suppose a test measures K binary attributes, producing 2^K latent classes. Let $\boldsymbol{\alpha}_c = (\alpha_{c1}, \dots, \alpha_{cK})^T$ denote the attribute profile vector for latent class c , where $c = 1, \dots, 2^K$. Element $\alpha_{ck} = 1$, if attribute k is mastered by individuals in latent class c , and $\alpha_{ck} = 0$, if attribute k is not mastered. The sG-DINA model (Ma & de la Torre, 2016) assumes that item $j \in \{1, \dots, J\}$ involves H_j tasks that need to be solved sequentially, and that students obtain a score of 0 if they fail the first task, a score of h ($0 < h < H_j$) if they perform the first h tasks successfully, but fail task $h + 1$, and a score of H_j if they perform all tasks successfully. The probability of individuals in latent class c performing task h correctly given that task $h - 1$ has been completed successfully is referred to as the processing function (Samejima, 1997) and denoted as $s_{jh}(\boldsymbol{\alpha}_c)$.

Given that different tasks of item j may involve different attributes, a binary q -vector \mathbf{q}_{jh} can be used to specify whether each attribute is measured by task h of item j , where element $q_{jkh} = 1$ if attribute k is measured, and $q_{jkh} = 0$ if not. A collection of \mathbf{q}_{jh} produces a category level Q -matrix with $\sum_{j=1}^J H_j$ rows. For task h of item j , let $\boldsymbol{\alpha}_{ljh}^*$ be the reduced attribute profile consisting of the required attributes for this task only, where $l = 1, \dots, 2^{K_{jh}^*}$ when the first K_{jh}^* attributes are assumed to be required. Note that 2^K latent classes can be collapsed into $2^{K_{jh}^*}$ latent groups for category h of item j , and $s_{jh}(\boldsymbol{\alpha}_c) = s(\boldsymbol{\alpha}_{ljh}^*)$ when latent class c is collapsed into latent group l . The sG-DINA model defines the processing function using the G-DINA model (de la Torre, 2011) as in

$$s(\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \phi_{jkh} \alpha_{lk} + \sum_{k'=k+1}^{K_{jh}^*} \sum_{k=1}^{K_{jh}^*-1} \phi_{jkhk'} \alpha_{lk} \alpha_{lk'} + \dots + \phi_{jh12\dots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk}, \quad (1)$$

where $\boldsymbol{\phi}_{jh} = (\phi_{jh0}, \dots, \phi_{jh12\dots K_{jh}^*})^T$ is a vector of parameters involved in category h of item j and $\boldsymbol{\Phi}$ is used to denote a vector of all parameters involved in the measurement model. By setting appropriate constraints as in de la Torre (2011), the DINA, DINO, and A -CDM can also be used as the processing function, if necessary, for different categories within a single item. Specifically, the sequential DINA (sDINA) model is obtained when all main effects and interaction terms except the highest-order interaction are set to be 0:

$$s(\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \phi_{jh12\dots K_{jh}^*} \prod_{k=1}^{K_{jh}^*} \alpha_{lk}. \quad (2)$$

The sequential DINO (sDINO) model is given by

$$s(\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \phi_{jkh} \alpha_{lk}, \quad (3)$$

where $\phi_{jkh} = -\phi_{jkhk'} = \dots = (-1)^{K_{jh}^*+1} \phi_{jh12\dots K_{jh}^*}$, for $k = 1, \dots, K_{jh}^*$, $k' = 1, \dots, K_{jh}^* - 1$, and $k'' > k', \dots, K_{jh}^*$. The sequential A -CDM (sA-CDM) is the constrained identity-link G-DINA model without any interaction terms. It can be formulated as follows:

$$s(\boldsymbol{\alpha}_{ljh}^*) = \phi_{jh0} + \sum_{k=1}^{K_{jh}^*} \phi_{jkh} \alpha_{lk}. \quad (4)$$

Limited-Information Measures

Let $\boldsymbol{\pi} = \{\pi_x\}$ be a vector of length u containing the (population) probabilities of each response pattern \mathbf{x} , and $\mathbf{p} = \{p_x\}$ be the corresponding observed proportions from a sample of size N . Also, let $\hat{\boldsymbol{\pi}} = \{\pi_x(\hat{\boldsymbol{\gamma}})\}$ be the model-implied response pattern probabilities associated with each response pattern based on v parameter estimates $\hat{\boldsymbol{\gamma}}$. Bishop, Fienberg, and Holland (1975) showed that $\sqrt{N}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$. In addition, as shown by Maydeu-Olivares and Joe (2005), the asymptotic distribution of the residual vector $\mathbf{p} - \hat{\boldsymbol{\pi}}$ is normal with zero means and limiting covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}^T$, that is, $\sqrt{N}(\mathbf{p} - \hat{\boldsymbol{\pi}}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Delta} = \partial\boldsymbol{\pi}(\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}$ is the $u \times v$ Jacobian matrix and $\boldsymbol{\mathcal{I}} = \boldsymbol{\Delta}^T \text{diag}[\boldsymbol{\pi}(\boldsymbol{\gamma})]^{-1} \boldsymbol{\Delta}$ is the Fisher information matrix.

Let $\hat{\boldsymbol{\kappa}} = (\hat{\boldsymbol{\kappa}}_1^T, \hat{\boldsymbol{\kappa}}_2^T)^T$ be a vector of length $w = J(J+1)/2$ containing all univariate and bivariate expectations, where $\hat{\boldsymbol{\kappa}}_1$ and $\hat{\boldsymbol{\kappa}}_2$ have elements $\kappa_a(\hat{\boldsymbol{\gamma}}) = E[X_a]$ and $\kappa_{a,b}(\hat{\boldsymbol{\gamma}}) = E[X_a X_b]$, respectively. Also let $\mathbf{m} = (\mathbf{m}_1^T, \mathbf{m}_2^T)^T$, where \mathbf{m}_1 and \mathbf{m}_2 are the sample counterparts of $\hat{\boldsymbol{\kappa}}_1$ and $\hat{\boldsymbol{\kappa}}_2$,

respectively. It is straightforward to show that $\hat{\mathbf{k}}$ is a linear transformation of $\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$, that is, $\hat{\mathbf{k}} = \mathbf{L}\boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})$, where \mathbf{L} is a $w \times u$ matrix having full row rank. Likewise, $\mathbf{m} = \mathbf{L}\mathbf{p}$. It is clear that $\mathbf{m} - \hat{\mathbf{k}} = \mathbf{L}[\mathbf{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\gamma}})]$ is also normally distributed, $\sqrt{N}(\mathbf{m} - \hat{\mathbf{k}}) \xrightarrow{d} \mathcal{N}_w(\mathbf{0}, \boldsymbol{\Xi})$, where $\boldsymbol{\Xi} = \mathbf{L}\boldsymbol{\Sigma}\mathbf{L}^T = \mathbf{L}[\boldsymbol{\Gamma} - \boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}^T]\mathbf{L}^T = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^T - \mathbf{L}\boldsymbol{\Delta}\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}^T\mathbf{L}^T$. Denote $\boldsymbol{\Gamma}_\kappa = \mathbf{L}\boldsymbol{\Gamma}\mathbf{L}^T$ and $\boldsymbol{\Delta}_\kappa = \mathbf{L}\boldsymbol{\Delta}$. $\boldsymbol{\Xi} = \boldsymbol{\Gamma}_\kappa - \boldsymbol{\Delta}_\kappa\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{\Delta}_\kappa^T$. Let $\bar{\boldsymbol{\Delta}}_\kappa$ be an $w \times (w - \nu)$ orthogonal complement of $\boldsymbol{\Delta}_\kappa$ so that $\bar{\boldsymbol{\Delta}}_\kappa^T\boldsymbol{\Delta}_\kappa = \mathbf{0}$. The $w - \nu$ dimensional vector $\mathbf{z}_\kappa = \sqrt{N}\bar{\boldsymbol{\Delta}}_\kappa^T(\mathbf{m} - \hat{\mathbf{k}})$ is normally distributed with asymptotic covariance matrix $\bar{\boldsymbol{\Delta}}_\kappa^T\boldsymbol{\Gamma}_\kappa\bar{\boldsymbol{\Delta}}_\kappa$. The M_{ord} statistic (Maydeu-Olivares & Joe, 2014) is the quadratic form

$$M_{\text{ord}} = \mathbf{z}_\kappa^T \left[\bar{\boldsymbol{\Delta}}_\kappa^T \boldsymbol{\Gamma}_\kappa \bar{\boldsymbol{\Delta}}_\kappa \right]^{-1} \mathbf{z}_\kappa = N(\mathbf{m} - \hat{\mathbf{k}})^T \mathbf{C}_\kappa (\mathbf{m} - \hat{\mathbf{k}}), \quad (5)$$

where $\mathbf{C}_\kappa = \bar{\boldsymbol{\Delta}}_\kappa [\bar{\boldsymbol{\Delta}}_\kappa^T \boldsymbol{\Gamma}_\kappa \bar{\boldsymbol{\Delta}}_\kappa]^{-1} \bar{\boldsymbol{\Delta}}_\kappa^T$. Under the null hypothesis, M_{ord} is approximately χ^2 distributed. The degrees of freedom are $w - \nu$, where $w = J(J + 1)/2$ and ν is the number of parameters. Both $\boldsymbol{\Gamma}_\kappa$ and $\boldsymbol{\Delta}_\kappa$ are evaluated at $\hat{\boldsymbol{\gamma}}$. Let $\boldsymbol{\gamma} = (\boldsymbol{\phi}^T, \boldsymbol{\rho}^T)^T$ consist of both measurement model parameters $\boldsymbol{\phi}$ and structural parameters $\boldsymbol{\rho}$, and the resulting M_{ord} is denoted as $M_{\text{ord}}^{\text{all}}$ because all model parameters are considered. The number of parameters in $\boldsymbol{\rho}$ increases exponentially with the number of attributes K , and therefore, when K is very large, the M_{ord} statistic may not be calculable. To address this issue, flexMIRT (Cai, 2017) ignores the structural parameter $\boldsymbol{\rho}$ when calculating the M_2 statistic. However, the impact of ignoring the structural parameters has not been documented. Therefore, M_{ord} statistic (referred to as $M_{\text{ord}}^{\text{item}}$) with $\boldsymbol{\gamma} = \boldsymbol{\phi}$ is also calculated. The calculation details of the M_{ord} statistics for the sG-DINA model are given in the Supplemental Appendix. In addition to the M_{ord} statistics, the SRMSR (Maydeu-Olivares, 2013) can also be calculated as

$$\text{SRMSR} = \sqrt{\sum_{a < b} \frac{(r_{ab} - \hat{\rho}_{ab})^2}{J(J-1)/2}}, \quad (6)$$

where r_{ab} and $\hat{\rho}_{ab}$ are observed and model-implied Pearson correlations, respectively, for items a and b . This index can be viewed as an average of correlation residuals for all item pairs.

Simulation Studies

In this section, two simulation studies were conducted to evaluate the viability of the $M_{\text{ord}}^{\text{all}}$ and $M_{\text{ord}}^{\text{item}}$ statistics. Study 1 investigated their performance when the Q-matrix was correctly specified, but the fitted measurement model may or may not conform to the underlying condensation rule; in contrast, Study 2 considered the conditions where the Q-matrix was mistakenly specified, but the fitted measurement model is in line with the underlying condensation rule. The factors manipulated are summarized in Table 1, and elaborated thereafter.

Study 1: The Model-Data Fit Measures Under Condensation Rule Misspecifications

Design. The number of items was fixed to $J = 30$, which has been considered in many previous simulation studies (e.g., Y. Liu, Tian, & Xin, 2016) and is also similar to the test length in real-world diagnostic assessments (e.g., Bradshaw, Izsák, Templin, & Jacobson, 2014). Sample sizes were 1,000 and 3,000, where the former is close to the median of sample sizes (i.e., 1,255) of 36 articles on CDM applications reviewed by Sessoms and Henson (2018) and the later represents a relatively large, but still realistic sample size as 30% of articles reviewed by Sessoms

Table 1. Summary of the Simulation Factors.

Factors	Study 1	Study 2
Sample size (<i>N</i>)	1,000, 3,000	
Test length (<i>J</i>)	30	
Number of attributes (<i>K</i>)	5	
Number of response categories (<i>RC</i>)	3, 4	
Generating attribute structure	Multivariate normal distribution	
Item quality	High, moderate, low	
Generating model	sG-DINA, sDINA, sDINO, sA-CDM	sG-DINA
Fitted model	sG-DINA, sDINA, sDINO, sA-CDM	sG-DINA
Proportion of misspecified <i>q</i> -entries	0%	5%

Note. DINA = deterministic input noisy “and” gate; DINO = deterministic input noisy “or” gate; CDM = cognitive diagnosis model; A-CDM = additive CDM; sG-DINA = sequential generalized DINA model; sDINA = sequential DINA; sDINO = sequential DINO; sA-CDM = sequential A-CDM.

and Henson (2018) had sample sizes greater than 2000. The number of response categories, which is identical for all items in a test, has two levels: $RC = 3$ or 4 with the maximum score being 2 or 3, respectively. The Q-matrix was simulated for each replication with constraints that (a) the maximum number of attributes required by each nonzero category is 2, (b) the number of categories measuring single and two attributes are equal, (c) each nonzero category measures at least one attribute, and (d) each attribute was measured by at least one item. Like Chiu, Douglas, and Li (2009), for individual i , latent traits $\theta_i = (\theta_{i1}, \dots, \theta_{iK})^T$ were first generated from a multivariate normal distribution with mean vector $\mathbf{0}_K$. Variances and covariances in the covariance matrix were set to 1 and 0.5, respectively. Then the k th element of attribute profile $\alpha_{ik} = 1$ if $\theta_{ik} \geq \Phi^{-1} k / (K + 1)$ and 0 otherwise. Data were simulated using the sDINA model, sDINO model, sA-CDM, and sG-DINA model. The quality of items had three levels with both $s(\alpha_{ljh}^* = \mathbf{0})$ and $1 - s(\alpha_{ljh}^* = \mathbf{1})$ being drawn from $U(0.05, 0.15)$, $U(0.15, 0.25)$, and $U(0.25, 0.35)$ for all categories of all items, representing high, moderate, and low quality, respectively. For the sA-CDM, main effects were constrained to be equal for each item indicating that all required attributes have the same contribution to the processing function as in Ma and de la Torre (2019). For the sG-DINA model, the success probabilities for individuals with attribute pattern α_{ljh}^* being neither $\mathbf{0}$ nor $\mathbf{1}$ were simulated randomly with the monotonic constraint that $s(\alpha_{ljh}^*) \geq s(\alpha_{l'jh}^*)$ if $\alpha_{ljh}^* \succ \alpha_{l'jh}^*$.

Note that whether the M_{ord} statistics can distinguish different sequential models rely on how similar or dissimilar they are. Ma, Iaconangelo, and de la Torre (2016) examined the similarity among several dichotomous CDMs, but they mainly focused on additive models with different link functions and did not consider the impact of the quality of items. In this study, the dissimilarity between a true model and an approximating model is defined based on the Kullback–Leibler divergence (Chang & Ying, 1996; Xu, Chang, & Douglas, 2003), where a small value indicates that the approximating model can mimic the true model well and a large value indicates that the approximating model cannot mimic the true model well. Definition of the dissimilarity index and details about the dissimilarity analysis can be found online in the Supplemental Appendix.

Figure 1 gives the boxplot of the dissimilarity among different models based on 500 replications under varied conditions. Several findings can be observed. First, under all conditions, the dissimilarity between a model and itself was 0, and the sG-DINA model mimicked the models it subsumes perfectly. Second, the additive processing function represents a condensation rule

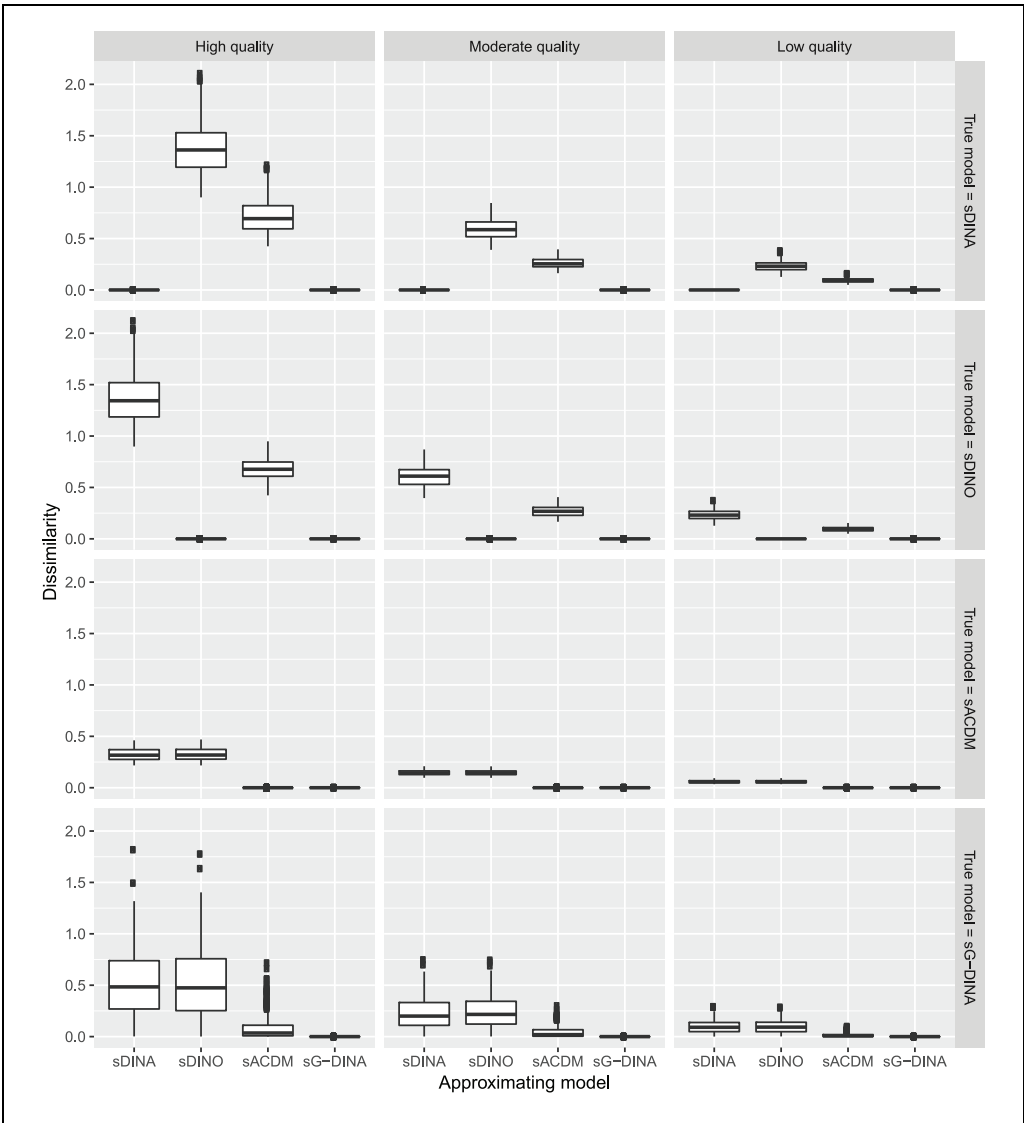


Figure 1. Dissimilarity among different processing functions.

Note. DINA = deterministic input noisy “and” gate; DINO = deterministic input noisy “or” gate; CDM = cognitive diagnosis model; sG-DINA = sequential generalized DINA model; sDINA = sequential DINA; sDINO = sequential DINO; sA-CDM = sequential additive CDM.

between the conjunctive and disjunctive rules. Specifically, the sA-CDM mimicked the sDINA model better than the sDINO model did, and mimicked the sDINO model better than the sDINA model did. Third, the sA-CDM can mimic the sDINA and sDINO models similarly well. Fourth, it is difficult for the sDINA and sDINO models to mimic the sG-DINA model, whereas the sA-CDM can mimic the sG-DINA model quite well under many replications. Last but not least, the worse the quality of items became, the better one model could mimic another. This suggests that the quality of items and the dissimilarity of models are confounded. In other words, although

the label of “item quality” was used, it also represents the magnitude of dissimilarity to some extent.

Under each condition, 500 data sets were generated and fitted using the sDINA, sDINO, sA-CDM, and sG-DINA models. The M_{ord}^{all} and M_{ord}^{item} statistics were calculated for each fitted model. All analyses were conducted using the GDINA R package (Ma & de la Torre, 2018). The Type I error rate of the M_{ord} statistic is calculated as the proportion of the generating model that is mistakenly flagged as a misfit model over all replications under each condition. Also, the correct detection rate is defined as the proportion of a misspecified condensation rule or Q-matrix that was correctly flagged over all replications.

Results. Table 2 gives the Type I error rates of the M_{ord}^{all} and M_{ord}^{item} statistics. Note that with 500 replications, the Type I error rates have a 95% chance of falling in the interval of [.04, .06]. From Table 2, it can be observed that the M_{ord}^{all} statistic had well-calibrated Type I error rates under the .05 nominal level with only a few exceptions. Specifically, the M_{ord}^{all} statistic was slightly liberal for the sDINA model with four out of 12 Type I error rates between .06 and .075, but slightly conservative for the sG-DINA model with two Type I error rates between .03 and .04. In contrast, the M_{ord}^{item} statistic was more likely to produce underestimated Type I error rates, especially when items were of high quality. Two inflated Type I error rates of M_{ord}^{item} statistic were observed when items were of moderate and low quality. Sample size, the number of response categories, and item quality had little impact on the Type I error rates.

The correct detection rates for M_{ord}^{all} and M_{ord}^{item} statistics for data generated from the sDINA model are presented in Table 3. In the literature, a correct detection rate of .80 or higher is typically considered adequate, and .90 or higher excellent (e.g., de la Torre & Lee, 2013). When items were of moderate or high quality, both M_{ord}^{all} and M_{ord}^{item} statistics had excellent correct detection rates in rejecting the sDINO model, with only one exception, which occurred under $N=1,000$, $RC=4$, and moderate item quality condition. When items were of low quality, the M_{ord}^{all} statistic had low detection rates to reject the sDINO model. In contrast, the M_{ord}^{item} statistic had higher detection rates. For example, when $N=3,000$, $RC=3$, and items were of low quality, the correct detection rates of M_{ord}^{all} and M_{ord}^{item} statistics were 0.282 and 0.980, respectively. In addition, as shown in Table 3, the correct detection rates for the M_{ord}^{all} and M_{ord}^{item} statistics in rejecting the sA-CDM were excellent when items were of high quality and $N=3,000$, but dropped dramatically as N became smaller or item quality became less optimal. For example, when items were of low quality, the correct detection rates for the M_{ord}^{all} and M_{ord}^{item} statistics were below .082 and .104, respectively. Although M_{ord}^{item} statistic still tended to have higher correct detection rates in rejecting the sA-CDM than M_{ord}^{all} statistic under most conditions, the differences were less noticeable. Last, Table 3 also shows that the correct detection rates in rejecting the sG-DINA model for M_{ord}^{all} and M_{ord}^{item} statistics were very low (ranging from .024 to .074), which was consistent with the author’s expectation in that the generating model is subsumed by the sG-DINA model. This implies that the M_{ord} statistic is insensitive to model overfitting. Similar patterns were also observed when data were generated from other sequential models subsumed by the sG-DINA model.

When data were generated using the sDINO model, as presented in Table 4, the correct detection rates in rejecting the sDINA model for M_{ord}^{all} and M_{ord}^{item} statistics are adequate when items were of moderate or high quality (ranging from .866 to 1); but dropped considerably as item quality became poor, especially for the M_{ord}^{all} statistic. More specifically, when items were of low quality, the correct detection rates for the M_{ord}^{all} statistic ranged from .074 to .260; whereas the M_{ord}^{item} statistic outperformed the M_{ord}^{all} statistic with correct detection rates ranging from .266 to .960. In addition, from Table 4, the correct detection rates in rejecting the sA-CDM for M_{ord}^{all} and M_{ord}^{item} statistics were excellent when items were of high quality, but worsened substantially

Table 2. Type I Error Rates Under $\alpha = .05$.

Model	N	RC	High quality		Moderate quality		Low quality	
			M_{ord}^{all}	M_{ord}^{all}	M_{ord}^{all}	M_{ord}^{all}	M_{ord}^{all}	M_{ord}^{all}
sDINA	1,000	3	0.058	0.030	0.048	0.040	0.054	0.050
		4	0.066	0.038	0.060	0.044	0.062	0.062
	3,000	3	0.054	0.020	0.074	0.044	0.070	0.056
		4	0.058	0.040	0.040	0.036	0.044	0.046
sDINO	1,000	3	0.056	0.016	0.060	0.044	0.056	0.042
		4	0.054	0.026	0.052	0.052	0.048	0.054
	3,000	3	0.048	0.024	0.036	0.030	0.048	0.050
		4	0.048	0.032	0.050	0.034	0.050	0.048
sA-CDM	1,000	3	0.054	0.042	0.034	0.024	0.064	0.048
		4	0.048	0.046	0.060	0.050	0.046	0.046
	3,000	3	0.054	0.034	0.050	0.042	0.052	0.044
		4	0.052	0.038	0.050	0.044	0.048	0.052
sG-DINA	1,000	3	0.044	0.034	0.058	0.050	0.056	0.054
		4	0.044	0.038	0.044	0.038	0.038	0.040
	3,000	3	0.034	0.030	0.046	0.046	0.036	0.038
		4	0.040	0.032	0.058	0.068	0.040	0.048

Note. DINA = deterministic input noisy "and" gate; DINO = deterministic input noisy "or" gate; CDM = cognitive diagnosis model; sG-DINA = sequential generalized DINA model; sDINA = sequential DINA; sDINO = sequential DINO; sA-CDM = sequential additive CDM.

Table 3. Correct Detection Rates Under $\alpha = .05$: sDINA-Generated Data.

Model	N	RC	High quality		Moderate quality		Low quality	
			M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}
sDINO	1,000	3	1.000	1.000	0.902	1.000	0.100	0.546
		4	1.000	1.000	0.882	0.990	0.092	0.300
	3,000	3	1.000	1.000	0.996	1.000	0.282	0.980
		4	1.000	1.000	0.996	1.000	0.188	0.790
sA-CDM	1,000	3	0.798	0.810	0.092	0.118	0.064	0.046
		4	0.402	0.418	0.026	0.030	0.056	0.048
	3,000	3	1.000	1.000	0.470	0.568	0.082	0.104
		4	0.926	0.942	0.202	0.236	0.048	0.062
sG-DINA	1,000	3	0.064	0.024	0.068	0.056	0.068	0.048
		4	0.046	0.036	0.048	0.052	0.074	0.086
	3,000	3	0.048	0.026	0.052	0.034	0.058	0.060
		4	0.058	0.028	0.054	0.042	0.038	0.054

Note. DINA = deterministic input noisy "and" gate; DINO = deterministic input noisy "or" gate; CDM = cognitive diagnosis model; sG-DINA = sequential generalized DINA model; sDINO = sequential DINO; sA-CDM = sequential additive CDM.

under other studied conditions. For example, when items were of moderate quality, the detection rates ranged from .194 to .670.

Table 5 gives the correct detection rates for data generated from the sA-CDM. The M_{ord}^{all} statistic had adequate correct detection rates in rejecting the sDINA and sDINO models only when items were of high quality and $N = 3,000$. The M_{ord}^{item} statistic, however, had higher correct

Table 4. Correct Detection Rates Under $\alpha = .05$: sDINO-Generated Data.

Model	N	RC	High quality		Moderate quality		Low quality	
			M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}
sDINA	1,000	3	1.000	1.000	0.912	1.000	0.114	0.540
		4	1.000	1.000	0.866	0.988	0.074	0.266
	3,000	3	1.000	1.000	1.000	1.000	0.260	0.960
		4	1.000	1.000	1.000	1.000	0.182	0.728
sA-CDM	1,000	3	0.958	0.962	0.248	0.228	0.064	0.068
		4	0.932	0.926	0.194	0.200	0.080	0.070
	3,000	3	0.998	1.000	0.608	0.670	0.110	0.102
		4	1.000	1.000	0.474	0.486	0.070	0.048
sG-DINA	1,000	3	0.042	0.024	0.058	0.046	0.058	0.038
		4	0.058	0.022	0.056	0.040	0.054	0.062
	3,000	3	0.050	0.024	0.054	0.038	0.052	0.040
		4	0.042	0.020	0.042	0.036	0.054	0.064

Note. DINA = deterministic input noisy “and” gate; DINO = deterministic input noisy “or” gate; CDM = cognitive diagnosis model; sG-DINA = sequential generalized DINA model; sDINA = sequential DINA; sA-CDM = sequential additive CDM.

Table 5. Correct Detection Rates Under $\alpha = .05$: sA-CDM-Generated Data.

Model	N	RC	High quality		Moderate quality		Low quality	
			M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}
sDINA	1,000	3	0.454	1.000	0.028	0.806	0.044	0.142
		4	0.496	0.992	0.042	0.426	0.038	0.054
	3,000	3	1.000	1.000	0.162	0.998	0.052	0.432
		4	0.968	1.000	0.138	0.936	0.060	0.184
sDINO	1,000	3	0.488	1.000	0.022	0.856	0.032	0.134
		4	0.472	1.000	0.034	0.548	0.034	0.064
	3,000	3	0.974	1.000	0.202	1.000	0.040	0.572
		4	0.982	1.000	0.128	0.966	0.048	0.244
sG-DINA	1,000	3	0.048	0.036	0.040	0.036	0.050	0.044
		4	0.048	0.046	0.050	0.058	0.048	0.054
	3,000	3	0.058	0.026	0.058	0.050	0.058	0.048
		4	0.058	0.038	0.040	0.046	0.046	0.048

Note. DINA = deterministic input noisy “and” gate; DINO = deterministic input noisy “or” gate; sG-DINA = sequential generalized DINA model; sDINA = sequential DINA; sDINO = sequential DINO.

detection rates than the M_{ord}^{all} statistic under all conditions in rejecting the sDINA and sDINO models. Specifically, excellent correct detection rates for the M_{ord}^{item} statistic were observed when items were of high quality or $N = 3,000$, and adequate correction rates were observed when $N = 1,000$, $RC = 3$, and items were of moderate quality. However, when items were of low quality, both M_{ord}^{all} and M_{ord}^{item} statistics had low correct detection rates in rejecting the sDINA and sDINO models.

Table 6 gives the correct detection rates for data generated using the sG-DINA model. It can be found that both M_{ord}^{all} and M_{ord}^{item} statistics had adequate or better detection rates in rejecting the sDINA and sDINO models under favorable conditions (i.e., larger N , smaller RC , better item quality). However, their detection rates dropped to below adequate level under

Table 6. Correct Detection Rates Under $\alpha = .05$: sG-DINA-Generated Data.

Model	N	RC	High quality		Moderate quality		Low quality	
			M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}	M_{ord}^{all}	M_{ord}^{item}
sDINA	1,000	3	1.000	1.000	0.850	0.976	0.172	0.324
		4	0.986	0.998	0.452	0.778	0.086	0.166
	3,000	3	1.000	1.000	0.996	0.998	0.544	0.868
		4	1.000	1.000	0.924	0.992	0.230	0.532
sDINO	1,000	3	1.000	1.000	0.860	0.986	0.158	0.348
		4	1.000	1.000	0.612	0.906	0.106	0.186
	3,000	3	1.000	1.000	0.990	1.000	0.582	0.908
		4	1.000	1.000	0.962	1.000	0.292	0.650
sA-CDM	1,000	3	0.076	0.064	0.054	0.056	0.056	0.040
		4	0.062	0.056	0.028	0.042	0.060	0.064
	3,000	3	0.206	0.178	0.058	0.048	0.052	0.056
		4	0.176	0.138	0.054	0.062	0.052	0.048

Note. DINA = deterministic input noisy "and" gate; DINO = deterministic input noisy "or" gate; CDM = cognitive diagnosis model; sDINA = sequential DINA; sDINO = sequential DINO; sA-CDM = sequential additive CDM.

unfavorable conditions. A substantial improvement in the correct detection rates can be observed by using M_{ord}^{item} statistic instead of M_{ord}^{all} statistic under these unfavorable conditions in rejecting the sDINA and sDINO models. Last, under all studied conditions, two statistics performed similarly poorly in rejecting the sA-CDM with the correct detection rates ranging from .028 to .178, which suggests that M_{ord} statistics are insensitive to the omission of attribute interactions.

In addition to the M_{ord}^{all} and M_{ord}^{item} statistics, the properties of SRMSR under varied conditions were also investigated. As an absolute fit measure, the SRMSR is usually compared with a cut-off to indicate whether the model can fit the data adequately. The author attempts to find a cut-off ϵ such that it can be used to separate the true model from misspecified models or a model with $SRMSR < \epsilon$ produce relatively accurate person classifications as the major goal of CDM analyses is to classify students into latent classes. The accuracy of person classification is quantified by the proportion of correctly classified attribute vectors (PCV). Results are discussed below, but due to space limits, the scatter plots of PCV and SRMSR under different generating models are given online in the Supplemental Appendix.

When data were generated using the sDINA model, the sDINO model tend to produce larger SRMSRs and lower PCVs than the sDINA model. The sA-CDM produced larger SRMSRs when items were of high or moderate quality, but comparable SRMSRs when items were of low quality. The sG-DINA model produced similar SRMSRs and PCVs under all conditions. Taken together, SRMSR may be used to distinguish sDINO from sDINA model, but may not be effective to distinguish sA-CDM from sDINA model, especially when the quality of items was poor. It is apparent that SRMSR cannot be used to distinguish sG-DINA model from sDINA model.

In addition, compared with the number of response categories, the quality of items and sample size exerted a major influence on SRMSR. In particular, the SRMSR tended to be smaller as sample size increased for both true and misspecified models. When the quality of items worsened, the sDINA and sG-DINA models tended to produce slightly larger SRMSRs, but the sDINO and sA-CDM tended to produce smaller SRMSRs. More importantly, it is evident that no single cutoff ϵ of the SRMSR existed to differentiate the true model from misspecified models or to differentiate models that produced high PCVs from those with low PCVs.

Similar patterns can be observed when the generating model were other sequential models. Specifically, when the sDINO model was the true model, the SRMSR could differentiate the sDINO model from the sDINA model, but was less effective to differentiate it from the sA-CDM and sG-DINA models. When the sA-CDM was the true model, the SRMSR could distinguish the sA-CDM from sDINA and sDINO models, especially when items were of high or moderate quality but not from the sG-DINA model. When the sG-DINA model was the true model, the SRMSR can distinguish it from the sDINA and sDINO models especially when items were of high or moderate quality, but cannot distinguish it from sA-CDM usually. Also, regardless of the generating model, sample size and item quality had major impact on SRMSR and there was no single cutoff suitable for SRMSR to distinguish the true model from the misspecified models.

Study 2: The Model-Data Fit Measures Under Q-Matrix Misspecifications

Design. The goal of this simulation study is to examine the correct detection rates of the M_{ord}^{all} and M_{ord}^{item} statistics in detecting misspecifications in the Q-matrix. The settings of sample size, item quality, number of response categories, and generating attribute distribution were the same as the previous simulation study. The generating CDM considered in this simulation study is the sG-DINA model and the data were simulated in the same manner as in the previous study. For each replication, 2.5% 1 s and 2.5% 0 s in the Q-matrix were randomly selected and modified (i.e., $0 \rightarrow 1$, or $1 \rightarrow 0$) with the constraint that each row and each column of the misspecified Q-matrix contain at least one 0 and at least one 1. This yielded 5% balanced misspecifications in the Q-matrix, similar to de la Torre and Chiu (2016). Note that the proportion of misspecifications considered in this study was less than those in other studies on model fit evaluation (e.g., Y. Liu et al., 2016; Wang, Shu, Shang, & Xu, 2015) because a mild level of misspecifications is created, and higher detection rates could be expected if more elements in the Q-matrix are misspecified. The data were fitted using the sG-DINA model along with the misspecified Q-matrix, and the M_{ord}^{all} and M_{ord}^{item} statistics were calculated.

Results. Table 7 gives the correct detection rates of the M_{ord}^{all} and M_{ord}^{item} statistics, as well as the average SRMSRs and their 90% confidence intervals. The confidence intervals were calculated empirically from the 500 replications with the lower and upper bounds being the 5th and 95th percentile points, respectively. Several conclusions can be drawn from the results. First, the correct detection rates for both M_{ord}^{all} and M_{ord}^{item} statistics were higher under the conditions with larger sample size, better item quality, or fewer response categories. For example, when items were of high quality, both M_{ord}^{all} and M_{ord}^{item} statistics showed adequate or better correct detection rates (i.e., greater than .826). Note that only 5% elements in the Q-matrix were misspecified and it is expected that the correct detection rates would be improved when a larger proportion of elements are mistakenly specified. Second, across all conditions, the M_{ord}^{item} statistic outperformed the M_{ord}^{all} statistic with only one exception, where two statistics had the same correct detection rate. In addition to the M_{ord} statistics, SRMSRs varied considerably under different conditions. More specifically, SRMSR tend to be larger when items were of better quality and sample size was smaller. For example, when $N = 1,000$, items were of high quality and $RC = 3$, the 90% confidence interval for SRMSR is [.043, .102]; whereas when $N = 3,000$, items were of low quality and $RC = 3$, the 90% confidence interval for SRMSR is [.018, .027]. This implies that the SRMSR is not only a function of model-data misspecifications, but also of many other factors; hence, using a single cut-off of SRMSR under varied conditions to evaluate the magnitude of model-data misfit may not be appropriate.

Table 7. Correct Detection Rates of M_{ord} Statistics and SRMSR Under Misspecified Q-Matrix.

N	Item quality	RC	Correct detection rates		SRMSR	90%CI for SRMSR	
			M_{ord}^{all}	M_{ord}^{item}		LL	UL
1,000	High	2	0.950	0.954	0.071	0.043	0.102
		3	0.826	0.874	0.072	0.041	0.107
	Moderate	2	0.706	0.732	0.045	0.032	0.061
		3	0.384	0.540	0.046	0.034	0.061
	Low	2	0.208	0.236	0.033	0.029	0.037
		3	0.090	0.136	0.033	0.031	0.037
3,000	High	2	0.998	0.998	0.068	0.036	0.101
		3	0.938	0.960	0.067	0.038	0.099
	Moderate	2	0.924	0.952	0.038	0.024	0.054
		3	0.758	0.864	0.040	0.025	0.057
	Low	2	0.498	0.586	0.022	0.018	0.027
		3	0.188	0.312	0.022	0.019	0.027

Note. SRMSR = standardized root mean square residual; LL = lower limit; UL = upper limit.

Summary and Discussion

Assessing model-data fit has become a routine task in psychometric analyses to ensure the validity of inferences from the observed responses. This study systematically investigated the performance of two implementations of the M_{ord} statistic for ordinal response data under the sG-DINA model. Simulation studies showed that the M_{ord}^{all} statistic had better-calibrated Type I error rates than the M_{ord}^{item} statistic, which was more likely to be conservative, especially when items were of high quality. Neither statistic is sensitive in rejecting the sG-DINA model when data were simulated using CDMs it subsumed, with the correct detection rates being close to the nominal level. This is not unexpected in that based on the model dissimilarity analysis, the sG-DINA model has more parameters than the generating models and can mimic the generating models perfectly. These additional parameters would probably capture the idiosyncrasies in the data and yield an overfitted model. However, the overfitting is unlikely to be detected by the fit statistics that assess the magnitude of residuals in that the overfitted model produces similar, if not smaller, residuals as the true model. The Wald test and likelihood ratio test have been shown promising in comparing the sG-DINA model and models it subsumes (Ma & de la Torre, 2019), and thus may be used as a supplement to the M_{ord} statistics.

When the generating model was sDINA model, the M_{ord} statistics yielded higher correct detection rates in rejecting the sDINO model than the sA-CDM. Similarly, when the generating model was sDINO model, the M_{ord} statistics yielded higher correct detection rates in rejecting the sDINA model than the sA-CDM. This is caused by the fact that the sDINA and sDINO models represent two distinct condensation rules and that the sA-CDM is in-between, as shown in the boxplot of model dissimilarity in Figure 1. The dissimilarity among these three models explains why distinguishing the sDINA and sDINO models is easier than distinguishing them from the sA-CDM using the M_{ord} statistics.

In addition, when the generating model was sG-DINA model, the M_{ord} statistics had higher power to reject the sDINA and sDINO model than the sA-CDM. This is also in line with the findings from the study on dissimilarities among these models. From Figure 1, the sA-CDM mimicked the sG-DINA model better than the sDINA and sDINO models did, implying that distinguishing sA-CDM from sG-DINA model would be more challenging.

The dissimilarity analysis also reveals that the quality of items and model dissimilarity are confounded. Simulation studies showed that when items were of low quality, the M_{ord} statistics

had very low detection rates, which may be attributed to the fact that these models are too similar to differentiate. Note that when items were of low quality, the M_{ord} statistics also experienced considerable difficulties in detecting Q-matrix misspecifications. In addition to the conditions that involve items of poor quality, the correct detection rates of the M_{ord} statistics dropped substantially when sample size was small and number of response categories was large. The issue of low detection rates for M_{ord} statistic in detecting some types model misspecifications has been observed in other studies as well (e.g., Cai & Hansen, 2013). Therefore, the M_{ord} statistics need to be used with caution under these unfavorable conditions.

An interesting but perplexing finding is that the M_{ord}^{item} statistic tends to produce similar or higher correct detection rates than the M_{ord}^{all} statistic in detecting both model and Q-matrix misspecifications. It is unclear why this happens, and more studies are needed before it can safely be concluded that the M_{ord}^{item} statistic should be preferred, especially when it is noted the M_{ord}^{item} statistic tends to produce Type I error rates lower than the nominal level. Nevertheless, as shown in the real data analysis, which is presented online on the Supplemental Appendix due to the space limits, it is evident that the M_{ord}^{item} statistic is more likely to be calculable because it ignores population proportion parameters and thus has larger degrees of freedom than the M_{ord}^{all} statistic.

In addition to the M_{ord} statistics, which examines whether the model fits the data statistically, the author also investigates the performance of SRMSR, which estimates the magnitude of model-data misfit. Although researchers have used $SRMSR < .05$ as the cut-off for acceptable model-data fit in CDMs (R. George & Robitzsch, 2015; Jiang & Ma, 2018; R. Liu et al., 2018), this study showed that the ranges of SRMSRs varied considerably under different conditions for both true model and misspecified models and thus there is no one-size-fits-all cut-off for SRMSR. Given that the SRMSR is the average correlation residuals, a model with smaller SRMSR can be viewed as having better absolute fit on average. However, it is more informative to report the maximum correlation residuals similar to J. Chen, de la Torre, and Zhang (2013), which will allow researchers to identify how severe and where the worst fit is. In addition, it is possible to employ some resampling techniques to obtain the empirical distribution of the SRMSR, based on which, a formal hypothesis test can be performed to determine whether the obtained SRMSR is too large or not. However, the performance of resampling procedures needs to be further examined.

This study systematically documents the performance of M_{ord} statistics and SRMSR for the sG-DINA model, but it is not without limitations. First, the number of attributes and items were fixed in the simulation studies and future research may vary them. The attribute profiles were drawn from multivariate normal distribution with fixed correlations among attributes. Future research may vary the correlations or assume different distributions for generating attribute profiles. Also, like Ma and de la Torre (2019), when generating data using the sA-CDM, all required attributes were assumed to contribute equally, but future research could consider relaxing this constraint. In addition, this study only considered two types of misspecifications, namely, misspecified condensation rules and Q-matrix, and future studies may consider other causes of model-data misfit. For example, continuous latent variables may be treated as dichotomous ones and the number of attributes may be mistakenly specified. Furthermore, despite well-calibrated Type I error rates, the M_{ord} statistics were found insensitive to some types of misspecification. Future research may explore how to extend other measures, such as log odds ratio and transformed correlation (J. Chen et al., 2013), for ordinal responses. Also note that for dichotomous items, the M_{ord} statistic is equivalent to the M_2 statistic. Although the performance of M_2 statistic has been investigated for dichotomous response CDMs (Hansen et al., 2016; Y. Liu, Tian, & Xin, 2016), researchers did not take item quality or model dissimilarity into consideration. The current study suggests that the M_{ord} and M_2 statistics need to be used with caution when items were of poor quality. Last but not least, future research may explore how to integrate the M_{ord} statistics and SRMSR with Q-matrix validation procedures (e.g., de la Torre

& Chiu, 2016; Ma & de la Torre, 2020), item-level model comparison approaches (e.g., de la Torre & Lee, 2013) and item fit measures (e.g., J. Chen et al., 2013; Sorrel, Abad, Olea, de la Torre, & Barrada, 2017; Wang et al., 2015) to determine the most appropriate models with acceptable model-data fit.


Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Office for Research and Economic Development, The University of Alabama (Grant RG 14872).

ORCID iD

Wenchao Ma  <https://orcid.org/0000-0002-6763-0707>

Supplemental Material

Supplemental material for this article is available online.

References

- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice* (With the collaboration of R. J. Light & F. Mosteller). Cambridge, MA: MIT Press.
- Bradshaw, L., Izsák, A., Templin, J., & Jacobson, E. (2014). Diagnosing teachers' understandings of rational numbers: Building a multidimensional test within the diagnostic classification framework. *Educational Measurement: Issues and Practice, 33*, 2-14.
- Cai, L. (2017). flexMIRT: Flexible multilevel multidimensional item analysis and test scoring [Computer software] (Version 3.51). Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*, 245-276.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.
- Chen, F., Liu, Y., Xin, T., & Cui, Y. (2018). Applying the M_2 statistic to evaluate the fit of diagnostic classification models in the presence of attribute hierarchies. *Frontiers in Psychology, 9*, Article 1875.
- Chen, J., de la Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement, 50*, 123-140.
- Chiu, C.-Y., Douglas, J. A., & Li, X. (2009). Cluster analysis for cognitive diagnosis: Theory and applications. *Psychometrika, 74*, 633-665.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.
- de la Torre, J., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika, 81*, 253-273.
- de la Torre, J., & Lee, Y. S. (2013). Evaluating the Wald test for item-level comparison of saturated and reduced models in cognitive diagnosis. *Journal of Educational Measurement, 50*, 355-373.
- George, A. C., & Robitzsch, A. (2015). Cognitive diagnosis models in R: A didactic. *The Quantitative Methods for Psychology, 11*, 189-205.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301-321.
- Hansen, M., Cai, L., Monroe, S., & Li, Z. (2016). Limited-information goodness-of-fit testing of diagnostic classification item response models. *British Journal of Mathematical and Statistical Psychology, 69*, 225-252.

- Jiang, Z., & Ma, W. (2018). Integrating differential evolution optimization to cognitive diagnostic model estimation. *Frontiers in Psychology, 9*, Article 2142.
- Jurich, D. P. (2014). *Assessing model fit of multidimensional item response theory and diagnostic classification models using limited-information statistics* (Unpublished doctoral dissertation). James Madison University, Harrisonburg, VA.
- Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*, 357-383.
- Liu, R., & Jiang, Z. (2018). Diagnostic classification models for ordinal item responses. *Frontiers in Psychology, 9*, Article 2512.
- Liu, Y., Tian, W., & Xin, T. (2016). An application of M_2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics, 41*, 3-26.
- Liu, Y., Xin, T., Li, L., Tian, W., & Liu, X. (2016). An improved method for differential item functioning detection in cognitive diagnosis models: An application of Wald statistic based on observed information matrix. *Acta Psychologica Sinica, 48*, 588-598.
- Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology, 69*, 253-275.
- Ma, W., & de la Torre, J. (2018). GDINA: The generalized DINA model framework [Computer software] (Version 2.1). Retrieved from <https://CRAN.R-project.org/package=GDINA>
- Ma, W., & de la Torre, J. (2019). Category-level model selection for the sequential G-DINA model. *Journal of Educational and Behavioral Statistics, 44*, 45-77.
- Ma, W., & de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential G-DINA model. *British Journal of Mathematical and Statistical Psychology, 73*, 143-166.
- Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement, 40*, 200-217.
- Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives, 11*, 71-101.
- Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2 n contingency tables: A unified framework. *Journal of the American Statistical Association, 100*, 1009-1020.
- Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*, 713-732.
- Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*, 305-328.
- Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology, 61*, 331-360.
- Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85-100). New York, NY: Springer.
- Sessoms, J., & Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives, 16*, 1-17.
- Sorrel, M. A., Abad, F. J., Olea, J., de la Torre, J., & Barrada, J. R. (2017). Inferential item-fit evaluation in cognitive diagnosis modeling. *Applied Psychological Measurement, 41*, 614-631.
- Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.
- von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.
- Wang, C., Shu, Z., Shang, Z., & Xu, G. (2015). Assessing item-level fit for the DINA model. *Applied Psychological Measurement, 39*, 525-538.
- Xu, X., Chang, H., & Douglas, J. (2003, April). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, Chicago, IL.