*Article*

*Journal of Educational and Behavioral Statistics*
*2019, Vol. 44, No. 1, pp. 45–77*
*DOI: 10.3102/1076998618792484*
*Article reuse guidelines: sagepub.com/journals-permissions*

# Category-Level Model Selection for the Sequential G-DINA Model

**Wenchao Ma**
*College of Education, The University of Alabama*

**Jimmy de la Torre**
*Faculty of Education, The University of Hong Kong*

*Solving a constructed-response item usually requires successfully performing a sequence of tasks. Each task could involve different attributes, and those required attributes may be "condensed" in various ways to produce the responses. The sequential generalized deterministic input noisy "and" gate model is a general cognitive diagnosis model (CDM) for graded response items of this type. Although a host of dichotomous CDMs with different condensation rules can be used to parameterize the success probability of each task, specifying the most appropriate one remains challenging. If the CDM specified for each task is not in accordance with the underlying cognitive processes, the validity of the inference could be questionable. This study aims to evaluate whether several hypothesis tests, namely, the Wald test using various variance–covariance matrices, the likelihood ratio (LR) test, and the LR test using approximated parameters, can be used to select the appropriate CDMs for each task of graded response items. Simulation studies are conducted to examine the Type I error and power of the hypothesis tests under varied conditions. A data set from the Trends in International Mathematics and Science Study 2007 mathematics assessment is analyzed as an illustration.*

Keywords: *cognitive diagnosis; G-DINA; sequential CDM; polytomous data; model selection*

Cognitive diagnosis models (CDMs) have attracted considerable attention recently in the field of educational measurement. CDMs are multidimensional models with the intention of uncovering individuals' mastery profiles on a set of skills or attributes from their observed item responses. The attributes are typically, although not always, represented by binary latent variables with 1 for *mastery* and 0 for *nonmastery*.

To make inference about students' attribute profiles, a number of CDMs have been developed (for reviews, see DiBello, Roussos, & Stout, 2007). To understand these models, the condensation rule (Maris, 1999) is critical. The condensation rule defines the way that attributes interact to produce an observed item response. For example, based on a conjunctive condensation rule, the deterministic inputs, noisy “and” (DINA) gate (Haertel, 1989) model, assumes that individuals are expected to have low probabilities of performing an item correctly unless they master all required attributes. In contrast, based on the disjunctive condensation rule, the deterministic inputs, noisy “or” (DINO) gate (Templin & Henson, 2006) model assumes that mastering at least one required attribute could yield a high success probability. The *additive* CDM (A-CDM; de la Torre, 2011), which has an additive condensation rule, assumes that each required attribute contributes to the success probability independently and uniquely. Aside from these specific models, researchers have developed some general CDM frameworks. Examples are the generalized DINA (G-DINA; de la Torre, 2011) model, the log-linear CDM (LCDM; Henson, Templin, & Willse, 2009), and the general diagnostic model (von Davier, 2008). Note that the G-DINA model and LCDM consider all main effects of latent variables and all possible interactions among them. By setting appropriate constraints, specific models with conjunctive, disjunctive, or additive condensation rules can be obtained as special cases.

Specifying the condensation rule for each item is largely based on experts’ judgment and thus could be subjective. A misspecification in the condensation rules produces the use of inappropriate CDMs, which then results in a model-data misfit (Kunina-Habenicht, Rupp, & Wilhelm, 2012; Y. Liu, Tian, & Xin, 2016) and could call into question the validity of inferences. For example, Rojas, de la Torre, and Olea (2012) have shown that fitting the conjunctive model to the data generated from the disjunctive model, or vice versa, can lead to poor attribute estimation. With the development of the general CDMs, some may argue that the general models should be preferred to the reduced models, such as the DINA model, DINO model, and *A*-CDM because they can provide better model-data fit in terms of the likelihood. However, as noted by W. Ma, Iaconangelo, and de la Torre (2016), the reduced models may still be more appropriate for several reasons. For example, the reduced CDMs usually have more straightforward interpretations because of the corresponding condensation rules. In addition, due to fewer item parameters involved, the reduced models need a smaller sample for accurate parameter estimation. Lastly, W. Ma et al. (2016) have found that the appropriate reduced models can provide better person attribute estimation than the saturated models, especially when the sample size is small.

As emphasized by von Davier (2014), it is important to consider other alternatives prior to committing to using one particular model. A few studies along this line can be found in literature. For example, Chen, de la Torre, and Zhang (2013); Henson, Templin, and Willse (2009); and Sinharay and Almond (2007) evaluated and compared different models using Akaike’s (1974) information criterion (AIC), Schwarz’s (1978) Bayesian information criterion (BIC), and deviance information criterion (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) at the test level. A limitation of model comparison at the test level is that all items are typically assumed to conform to the same model, which, most likely, is not the case in practice.

At the item level, Henson et al. (2009) provided a way to determine the appropriate reduced model by visual inspection of the estimates of LCDM, and Sinharay and Almond (2007) checked the item fit plots generated from a residual analysis. In addition, de la Torre (2011) proposed to use the Wald test (Wald, 1943) to evaluate whether the reduced models subsumed by the G-DINA model can be used in place of the saturated G-DINA model without a significant loss in model-data fit. The Type I error and power of the Wald test for comparing the G-DINA model and the DINA model, DINO model and $A$-CDM were examined by de la Torre and Lee (2013), and the performance of the Wald test in comparing the G-DINA model with the logistic linear model (Maris, 1999) and reduced reparameterized unified model (Hartz, 2002) was later investigated by W. Ma et al. (2016). Although the model selected by the Wald test can provide better person classification, the Type I error rates of the Wald test are found to be inflated, especially when sample size is small or item quality is low (de la Torre & Lee, 2013; W. Ma et al., 2016). One possible cause is that the population proportion parameters were ignored when calculating the covariance matrix. Philipp, Strobl, de la Torre, and Zeileis (2018) found that the item parameter standard errors were underestimated when the population proportion parameters were ignored. Another possible cause is that all aforementioned studies estimated the covariance matrix using the outer product of the gradient (OPG) method, which has been shown less accurate for item response models when sample size is small (Chalmers, Pek, & Liu, 2017; Paek & Cai, 2014). One alternative way of estimating covariance matrix is developed by Louis (1982), who showed that the covariance matrix can be obtained using the code for expectation–maximization (EM) algorithm directly. Another alternative is the so-called sandwich-type estimator (White, 1982), which, unlike the OPG and Louis's estimators, is shown to be consistent under misspecified models (e.g., White, 1982; Yuan, Cheng, & Patton, 2014).

To date, research on model comparison, or condensation rule selection, mainly focuses on dichotomous responses data. In this article, we consider polytomously scored items that can be decomposed into a series of tasks. Items of this type are not uncommon, especially, in educational assessments. For example, when introducing the popular partial credit model (PCM), Masters (1982) took $\sqrt{7.5/0.3-16}$ as an example and identified three steps which "must be taken in order" (p. 155) to perform this item successfully. In addition, Hemker, van der Ark, and Sijtsma (2001) gave an example in an ability test for Dutch as a foreign language. Further, as noted by Masters (1982), the sequential step idea can also be applied to Likert-type scale for measuring psychological constructs. Although the PCM was originally developed to model items with a sequence of steps or tasks, it turns out to be suboptimal for this type of items because of the difficulty in interpreting item parameters (Verhelst & Verstralen, 2008; Tutz, 2016). Models belonging to the family of continuation ratio models, such as the sequential item response theory (IRT) model (Tutz, 2016; Verhelst, Glas, & de Vries, 1997) and the sequential G-DINA model (W. Ma & de la Torre, 2016), are more appropriate for items of this type. More importantly, unlike other polytomous response CDMs, the sequential G-DINA model allows us to specify the relation between each step and attributes and thus is particularly suitable for cognitively diagnostic assessments. However, the sequential G-DINA model parameterizes each of these tasks involved in a problem-solving sequence using the dichotomous G-DINA model (de la Torre, 2011), but different tasks may involve different condensation rules, and the saturated G-DINA model could be suboptimal due to the same reasons as in dichotomous responses.

This study aims to evaluate the performance of the Wald test using various covariance matrices and the likelihood ratio (LR) test in selecting appropriate condensation rules for the graded response items based on the sequential G-DINA model. Specifically, the OPG, Louis's, and sandwich covariance estimators are considered for the Wald test. It should be noted that this is not the first work investigating the sandwich-type covariance estimator in CDMs. Under the LCDM framework, Y. Liu, Xin, Andersson, and Tian (2018) evaluated the performance of covariance estimator based on the observed information matrix and sandwich-type estimator, which is based on the observed and OPG information matrices, using the dcminfo R package (Y. Liu & Xin, 2017). Based on these information matrices, Xin, Liu, Tian, and Li (2017) examined the performance of the Wald test for item-level model selection. However, the observed information matrix involves second derivatives of the log likelihood with respect to all model parameters, which are difficult to calculate analytically. In contrast, current study employs Louis's (1982) approach that tends to be easier to obtain when the EM algorithm is used for model estimation. In addition to the Wald test, this study considers the LR test for condensation rule selection to reduce the impact of covariance matrix estimations. Investigating condensation rule selection procedures for the sequential G-DINA model has the potential to advance the use of constructed-response items in cognitively diagnostic assessments. In addition, since the dichotomous G-DINA model is a special case of the sequential G-DINA model, the findings in this study have important implications about the performance of these procedures under the G-DINA model for dichotomous response data. The remainder of this article is laid out as follows. The second section provides an overview of the sequential G-DINA model. In the third section, we introduce how the LR test and Wald test are used for condensation rule selection. The fourth section describes in detail a simulation study for evaluating the Type I error and power of the LR test and Wald test under varied conditions. Then, a set of data from Trends in International Mathematics and Science Study (TIMSS) 2007 mathematics assessment was analyzed to illustrate how the Wald and LR tests can be used in practice. We conclude the sixth section with a brief summary of this study and a discussion of directions for future research.

## Overview of the Sequential G-DINA Model

Suppose a test measuring $K$ attributes has $J$ items. Also, suppose that item $j$ consists of $H_j$ tasks that need to be undertaken sequentially and yields $H_j + 1$ response categories (i.e., category 0, 1, $\ldots, H_j$). Specifically, a student gets a score of 0 if she or he fails the first task, and a score of $h$ ($h > 0$) if she or he performs the first $h$ tasks successfully, but fails task $h + 1$ if task $h$ is not the last task. Because task $h$ is related to response category $h$ directly, we use them interchangeably in this article. A binary q-vector $\boldsymbol{q}_{jh} = \{q_{jhk}\}$ is associated with task $h$ of item $j$, where element $q_{jhk} = 1$ if attribute $k$ is required by task $h$ and $q_{jhk} = 0$ otherwise. A collection of $\boldsymbol{q}_{jh}$ produces a category level Q-matrix, or Q$_\text{c}$-matrix (W. Ma & de la Torre, 2016), which is a $\sum_{j=1}^{J} H_j \times K$ binary matrix. If all items are scored dichotomously, the Q$_\text{c}$-matrix is equivalent to the traditional Q-matrix (Tatsuoka, 1983). Individuals can be grouped into $2^K$ latent classes because of the $K$ attributes involved in the assessment. Individuals in the same latent class have the same attribute pattern, which is denoted as $\boldsymbol{\alpha}_c = (\alpha_{c1}, \ldots, \alpha_{cK})'$ for latent class $c$, where $c = 1, \ldots, 2^K$. 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 sequential G-DINA model (W. Ma & de la Torre, 2016) assumes that solving an item involves undertaking a sequence of tasks and that individuals cannot complete a task unless they have already performed the previous task successfully. Let $\boldsymbol{Y}_i = \{Y_{ij}\}$ denote a $J$-dimensional random vector representing the responses of individual $i$ to $J$ items, and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_N)'$ be a random sample consisting of item responses of $N$ individuals, which are assumed to be independent and identically distributed. The probability of performing task $h$ correctly given that task $h - 1$ has been completed successfully is referred to as the processing function (W. Ma & de la Torre, 2016) and denoted as $s_{jh}(\boldsymbol{\alpha}_c) = P(Y_{ij} \geq h | Y_{ij} \geq h - 1, \boldsymbol{\alpha}_c)$. The conditional probability of obtaining score $h$ on item $j$ can be written as

$$P(Y_{ij} = h; \boldsymbol{\alpha}_c) = \left[1 - s_{j,h+1}(\boldsymbol{\alpha}_c)\right] \prod_{y=0}^{h} s_{jy}(\boldsymbol{\alpha}_c), \tag{1}$$

where $s_{j0}(\boldsymbol{\alpha}_c) \equiv 1$ and $s_{j,H_j+1}(\boldsymbol{\alpha}_c) \equiv 0$.

For task $h$ of item $j$, let $\boldsymbol{\alpha}^*_{ljh}$ be the reduced attribute pattern consisting of the required attributes for this task only. Without loss of generality, the first $K^*_{jh}$ attributes are assumed to be required, that is, $l = 1, \cdots, 2^{K^*_{jh}}$. If not all $K$ attributes are needed for this task, $2^K$ latent classes can be collapsed into $2^{K^*_{jh}}$ latent groups in that some latent classes have the same success probability. More formally, we can define a set of many-to-one mappings, $M_{jh}$, so that $\boldsymbol{\alpha}^*_{ljh} = M_{jh}(\boldsymbol{\alpha}_c)$ if latent class $c$ is collapsed into latent group $l$. Denote $C_{ljh} = \{\boldsymbol{\alpha}_c : M_{jh}(\boldsymbol{\alpha}_c) = \boldsymbol{\alpha}^*_{ljh}\}$, and if $\boldsymbol{\alpha}_c \in C_{ljh}$, we have $s_{jh}(\boldsymbol{\alpha}_c) = s(\boldsymbol{\alpha}^*_{ljh})$. Note that the subscripts for the processing function with reduced attribute patterns have been dropped for simplicity. The processing function can be expressed as

$$g[s(\boldsymbol{\alpha}^*_{ljh})] = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{lk} + \sum_{k'=k+1}^{K^*_{jh}} \sum_{k=1}^{K^*_{jh}-1} \phi_{jhkk'}\alpha_{lk}\alpha_{lk'} + \ldots + \phi_{jh12\cdots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk}, \tag{2}$$

where $g[\cdot]$ is the identity, log, or logit link function. By setting appropriate constraints to the identity link model as in de la Torre (2011), the DINA model, DINO model, or *A*-CDM can be used as the processing function, and different models can be used at different steps within a single item. Specifically, the DINA model is obtained when all main effects and interaction terms except the highest order interaction are set to be 0:

$$s(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \phi_{jh12\cdots K^*_{jh}} \prod_{k=1}^{K^*_{jh}} \alpha_{lk}. \tag{3}$$

The processing function based on the DINO model is given by

$$s(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \phi_{jhk}\alpha_{lk}, \tag{4}$$

where $\phi_{jhk} = -\phi_{jhk'k''} = \cdots = (-1)^{K^*_{jh}+1}\phi_{jh12\cdots K^*_{jh}}$, for $k = 1, \cdots, K^*_{jh}$, $k' = 1, \cdots, K^*_{jh} - 1$, and $k'' > k', \cdots, K^*_{jh}$. The *A*-CDM processing function is the constrained identity-link G-DINA model without the interaction terms. It can be formulated as

$$s(\boldsymbol{\alpha}^*_{ljh}) = \phi_{jh0} + \sum_{k=1}^{K^*_{jh}} \phi_{jhk}\alpha_{lk}. \tag{5}$$

## Category-Level Model Comparison

If a response category only requires one attribute, the G-DINA model and other reduced CDMs (e.g., DINA model, DINO model, and *A*-CDM) are not distinguishable, which implies that all condensation rules are equivalent. Therefore, model comparison is only necessary for categories requiring two or more attributes, which are referred to as multiattribute categories.

### *LR Test*

Let $\boldsymbol{s}_j = \{\boldsymbol{s}_{jh}\}$ be a vector of processing functions of all categories of item $j$, where $\boldsymbol{s}_{jh} = \{s(\boldsymbol{\alpha}^*_{ljh})\}$ and $\boldsymbol{s} = \{\boldsymbol{s}_j\}$ is a vector of all processing functions (i.e., item parameters) for the saturated sequential G-DINA model. Also, let $\boldsymbol{\pi} = \{\pi_c | c = 2, 3, \ldots, 2^K\}$ be free latent class proportion parameters and $\pi_1 = 1 - \sum_{c=2}^{2^K} \pi_c$. Let $\boldsymbol{\psi} = (\boldsymbol{s}', \boldsymbol{\pi}')'$ denotes a vector of all parameters involved in the model. The log likelihood of response vector $\boldsymbol{Y}_i$ for individual $i$ is

$$\ell(\boldsymbol{\psi}; \boldsymbol{Y}_i) = \log \sum_{c=1}^{2^K} \pi_c f(\boldsymbol{Y}_i; \boldsymbol{\alpha}_c), \tag{6}$$

where $f(\cdot)$ denotes a probability density function. The log likelihood of responses for all individuals is $\ell(\boldsymbol{\psi}; \boldsymbol{Y}) = \sum_{i=1}^{N} \ell(\boldsymbol{\psi}; \boldsymbol{Y}_i)$. The LR test has been widely used to compare two nested models: a compact model and an augmented model. Let $\ell(\boldsymbol{\psi}_C; \boldsymbol{Y})$ and $\ell(\boldsymbol{\psi}_A; \boldsymbol{Y})$ be the log likelihood of the compact and augmented models, respectively. The LR statistic can be written as $\text{LR} = -2[\ell(\boldsymbol{\psi}_C; \boldsymbol{Y}) - \ell(\boldsymbol{\psi}_A; \boldsymbol{Y})]$, which follows a $\chi^2$ distribution with the degrees of freedom equal to the difference in the number of parameters estimated for the two models.

In this study, the LR test is conducted category by category and item by item. For the augmented model, the G-DINA model is used as the processing functions for all categories of all items, whereas for the compact model, the G-DINA model is used as the processing functions for all categories except the studied category, for which a reduced model is used as the processing function. The augmented model only needs to be calibrated once. Given that the reduced models to be tested for each category in this study include the DINA model, DINO model, and A-CDM, the data need to be calibrated $1 + 3\sum_{j=1}^{J} \sum_{h=1}^{H_j} I(K_{jh}^* > 1)$ times.

The LR test could be time-consuming, and thus we also consider an EM-based approximation, which is referred to as two-step LR test, similar to the implementation in Sorrel, de la Torre, Abad, and Olea (2017). Specifically, the processing functions under a reduced model are estimated using a one-step EM algorithm based on the estimates under the G-DINA processing functions directly without recalibrating the data. When the DINA or DINO model is used as the processing function, some reduced latent groups are equivalent in that they have the same processing function. We use a vector of length $2^{K_{jh}^*}$, $\boldsymbol{\omega}_{jh}$, to denote the equivalent reduced latent groups for category $h$ of item $j$, where $\omega_{ljh} = g$ if $\boldsymbol{\alpha}_{ljh}^*$ is in the $g$th set of the equivalent reduced latent groups. For example, suppose $\boldsymbol{\alpha}_{jh}^* = \{00, 10, 01, 11\}$, then we have $\boldsymbol{\omega}_{jh} = (1, 2, 3, 4)$ for the G-DINA processing function, $\boldsymbol{\omega}_{jh} = (1, 1, 1, 2)$ for the DINA processing function, and $\boldsymbol{\omega}_{jh} = (1, 2, 2, 2)$ for the DINO processing function. The maximum likelihood estimate of $s(\boldsymbol{\alpha}_{ljh}^*)$ when $h \geq 1$ is given by

$$\hat{s}(\boldsymbol{\alpha}_{ljh}^*) = \frac{R_h^+(\boldsymbol{\alpha}_{ljh}^*)}{R_{h-1}^+(\boldsymbol{\alpha}_{ljh}^*)}, \tag{7}$$

where $R_h^+(\boldsymbol{\alpha}_{ljh}^*)$ is the expected number of examinees in reduced latent group $l$ and other equivalent groups getting at least a score of $h$ and can be calculated as

$$R_h^+(\boldsymbol{\alpha}_{ljh}^*) = \sum_{i=1}^{N} \sum_{\{l':\omega_{l'jh}=\omega_{ljh}\}} P(\boldsymbol{\alpha}_{l'jh}^*; \boldsymbol{Y}_i) I(Y_{ij} \geq h). \tag{8}$$

When the processing function is the $A$-CDM, the following log-likelihood function is maximized for a studied category while the processing functions of other categories are hold constant,

$$\sum_{c=1}^{2^K} \sum_{h=0}^{H_j} \bar{r}_{jhc} \log[P(Y_{ij} = h; \boldsymbol{\alpha}_c)], \tag{9}$$

where $\bar{r}_{jhc} = \sum_{i=1}^{N} P(\boldsymbol{\alpha}_c; \boldsymbol{Y}_i) I(Y_{ij} = h)$. Note that $P(\boldsymbol{\alpha}_{l'jh}^*; \boldsymbol{Y}_i)$ and $P(\boldsymbol{\alpha}_c; \boldsymbol{Y}_i)$ are calculated based on the augmented model.

## *The Wald Test*

To use the Wald test to examine whether a reduced model can be used in place of the G-DINA model as the processing function for a multiattribute category, category $h$ of item $j$, a $(2^{K_{jh}^*} - m) \times 2^{K_{jh}^*}$ restriction matrix $\boldsymbol{R}$ needs to be set up so that under the null hypothesis, $\boldsymbol{R} \times \boldsymbol{s}_{jh} = \boldsymbol{0}$, where $m$ is the number of parameters involved in this category when a reduced CDM is used as the processing function, and $s_{jh}$ is a vector consisting of the processing functions for category $h$ when the G-DINA model is used. For example, when $K_{jh}^* = 3$, the null hypothesis for the $A$-CDM is

$$\begin{bmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} s(000) \\ s(100) \\ s(010) \\ s(001) \\ s(110) \\ s(101) \\ s(011) \\ s(111) \end{bmatrix} = \mathbf{0}.$$

Examples of the restriction matrices for the DINA and DINO models can be found in de la Torre and Lee (2013). The Wald statistic can be calculated as

$$W = [\boldsymbol{R} \times \hat{\boldsymbol{s}}_{jh}]'[\boldsymbol{R} \times \boldsymbol{V}(\hat{\boldsymbol{s}}_{jh}) \times \boldsymbol{R}']^{-1}[\boldsymbol{R} \times \hat{\boldsymbol{s}}_{jh}], \tag{10}$$

where $\boldsymbol{V}(\hat{\boldsymbol{s}}_{jh})$ is the covariance matrix of the processing functions for category $h$ of item $j$. The Wald statistic $W$ is asymptotically $\chi^2$ distributed with $2^{K_{jh}^*} - m$ degrees of freedom.

An accurate estimation of the covariance matrix of the processing functions is critical for the Wald test. The covariance matrix of model parameters can be obtained by inverting the observed information matrix, which is defined as

$$\mathcal{I}(\boldsymbol{\psi}) = -\sum_{i=1}^{N}\left[\frac{\partial^2 \ell(\boldsymbol{\psi}; \boldsymbol{Y}_i)}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right]. \tag{11}$$

The observed information as defined in Equation 11 is typically hard to evaluate because of the second-order partial derivatives of the log-likelihood function; and therefore, in practice, it is often approximated by

$$\mathcal{I}_{\text{OPG}}(\widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{N}[\boldsymbol{S}(\boldsymbol{Y}_i; \widehat{\boldsymbol{\psi}})\boldsymbol{S}'(\boldsymbol{Y}_i; \widehat{\boldsymbol{\psi}})]. \tag{12}$$

where, with details given in Appendix A,

$$\boldsymbol{S}(\boldsymbol{Y}_i; \widehat{\boldsymbol{\psi}}) = \frac{\partial \log f(\boldsymbol{Y}_i; \widehat{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}}. \tag{13}$$

The information matrix $\mathcal{I}_{\text{OPG}}(\widehat{\boldsymbol{\psi}})$ is referred to as the OPG form by White (1982) and has been widely used and systematically studied in CDM context (e.g., Philipp, Strobl, de la Torre, & Zeileis, 2018) partially because of its simplicity of implementation.

In addition to the OPG approximation, Louis (1982) gives an expression to estimate the observed information matrix using the complete and missing information matrices from the EM algorithm directly. Let $\boldsymbol{a}_i \in \{\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_c, \ldots, \boldsymbol{\alpha}_{2^K}\}$ be a $K$-dimensional random vector representing the attribute pattern of individual $i$, and $\boldsymbol{X}_i = (\boldsymbol{Y}_i', \boldsymbol{a}_i')'$ be the "complete" data for individual $i$. As shown in Appendix B, the Louis's (1982) estimator can be written as

$$\begin{aligned}\mathcal{I}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}) = &\sum_{i=1}^{N}\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{B}(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i; \boldsymbol{Y}_i)\\ &-\sum_{i=1}^{N}\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{S}(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})\boldsymbol{S}'(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i; \boldsymbol{Y}_i)\\ &+\sum_{i=1}^{N}\left[\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{S}(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i; \boldsymbol{Y}_i)\right]\left[\sum_{a_{i1}=0}^{1}\cdots\sum_{\alpha_{iK}=0}^{1}\boldsymbol{S}'(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i; \boldsymbol{Y}_i)\right],\end{aligned} \tag{14}$$

where

$$\boldsymbol{S}(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}}) = \frac{\partial \log f(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}} \text{ and } \boldsymbol{B}(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}}) = \left[-\frac{\partial^2 \log f(\boldsymbol{X}_i; \widehat{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\right]. \tag{15}$$

The covariance matrix based on the OPG approximation and Louis's estimator can be written as $\boldsymbol{V}_{\text{OPG}}(\widehat{\boldsymbol{\psi}}) = \mathcal{I}_{\text{OPG}}^{-1}(\widehat{\boldsymbol{\psi}})$ and $\boldsymbol{V}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}) = \mathcal{I}_{\text{Louis}}^{-1}(\widehat{\boldsymbol{\psi}})$, respectively.

However, both $\boldsymbol{V}_{\text{OPG}}(\widehat{\boldsymbol{\psi}})$ and $\boldsymbol{V}_{\text{Louis}}(\widehat{\boldsymbol{\psi}})$ require a correctly specified model. When model is misspecified, only the sandwich estimator (Huber, 1967; White, 1982) is asymptotically consistent. Based on the information matrices estimated using the OPG and Louis's methods, as in Yuan, Cheng, and Patton (2014), the sandwich-type covariance matrix can be estimated as

$$\boldsymbol{V}_{\text{sw}}(\widehat{\boldsymbol{\psi}}) = \mathcal{I}^{-1}_{\text{Louis}}(\widehat{\boldsymbol{\psi}})\mathcal{I}_{\text{OPG}}(\widehat{\boldsymbol{\psi}})\mathcal{I}^{-1}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}). \tag{16}$$

Apart from various ways of estimating the covariance matrix, it is still questionable that what parameters need to be considered in the approximation of the observed information matrix for CDMs. In practice, only item parameters are typically of interest, and therefore, some previous studies (e.g., de la Torre, 2008) calculate the covariance matrix of item parameters by inverting the OPG information matrix for each item separately or, equivalently, by inverting a block diagonal information matrix for all items. Based on this, the Wald test has been shown to produce inflated Type I error rates under some conditions for model comparison and differential item functioning detection for dichotomous response data (de la Torre & Lee, 2013; Hou, de la Torre, & Nandakumar, 2014; W. Ma et al., 2016). Philipp et al. (2018) found that standard errors based on the complete OPG information matrix that considers both item and proportion parameters were more accurate. Despite the superiority, the complete information matrix could be too large to be manageable in that its size increases exponentially with the number of attributes. For example, when there are 15 attributes as in Lee, Park, and Taylan (2011), the complete information matrix is larger than 32,768 × 32,768, which may be problematic when calculating the inverse. Philipp et al. (2018) also examined the incomplete OPG information, which is a full matrix and considers only item parameters, and found that corresponding standard errors tended to be slightly underestimated. However, it is not clear whether the Wald test will be influenced. Therefore, in this study, we also consider estimating covariance matrix using only item parameters. The corresponding OPG, Louis, and sandwich-type estimators are denoted as $\boldsymbol{V}_{\text{OPG}}(\hat{\boldsymbol{s}}) = \mathcal{I}^{-1}_{\text{OPG}}(\hat{\boldsymbol{s}})$, $\boldsymbol{V}_{\text{Louis}}(\hat{\boldsymbol{s}}) = \mathcal{I}^{-1}_{\text{Louis}}(\hat{\boldsymbol{s}})$, and $\boldsymbol{V}_{\text{sw}}(\hat{\boldsymbol{s}}) = \mathcal{I}^{-1}_{\text{Louis}}(\hat{\boldsymbol{s}})\mathcal{I}_{\text{OPG}}(\hat{\boldsymbol{s}})\mathcal{I}^{-1}_{\text{Louis}}(\hat{\boldsymbol{s}})$, respectively.

## Simulation Study

The goal of this simulation study is to systematically evaluate the performance of the LR tests and the Wald test using different information matrices for category-level model selection in the context of the sequential G-DINA model. The Type I error and power of these statistical tests were examined under varied conditions.

### *Design*

The number of items and attributes were fixed to $J = 23$ and $K = 5$, respectively. The sample sizes were $N =$ 1,000, 2,000, and 4,000. The processing

TABLE 1.
$Q_c$-*Matrix for the Simulation Study*

| Item | Category | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | Item | Category | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 13 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 2 | 0 | 1 | 0 | 0 | 0 | 13 | 2 | 0 | 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 | 0 | 13 | 3 | 0 | 0 | 1 | 0 | 0 |
| 2 | 2 | 0 | 0 | 0 | 1 | 0 | 14 | 1 | 0 | 0 | 0 | 1 | 0 |
| 3 | 1 | 0 | 0 | 0 | 0 | 1 | 14 | 2 | 0 | 0 | 1 | 0 | 0 |
| 3 | 2 | 1 | 0 | 0 | 0 | 0 | 14 | 3 | 0 | 1 | 0 | 0 | 0 |
| 4 | 1 | 0 | 1 | 0 | 0 | 0 | 15 | 1 | 0 | 0 | 0 | 0 | 1 |
| 4 | 2 | 0 | 0 | 1 | 0 | 0 | 15 | 2 | 0 | 0 | 1 | 0 | 0 |
| 5 | 1 | 0 | 0 | 1 | 1 | 0 | 15 | 3 | 1 | 0 | 0 | 1 | 0 |
| 5 | 2 | 0 | 0 | 0 | 0 | 1 | 16 | 1 | 0 | 0 | 0 | 1 | 0 |
| 6 | 1 | 1 | 1 | 0 | 0 | 0 | 16 | 2 | 1 | 0 | 0 | 0 | 0 |
| 6 | 2 | 0 | 0 | 0 | 1 | 0 | 16 | 3 | 0 | 0 | 1 | 0 | 1 |
| 7 | 1 | 0 | 0 | 1 | 1 | 1 | 17 | 1 | 1 | 0 | 0 | 0 | 0 |
| 7 | 2 | 0 | 1 | 0 | 0 | 0 | 17 | 2 | 0 | 1 | 0 | 0 | 0 |
| 8 | 1 | 1 | 1 | 0 | 1 | 0 | 17 | 3 | 0 | 1 | 0 | 1 | 1 |
| 8 | 2 | 0 | 0 | 0 | 0 | 1 | 18 | 1 | 0 | 1 | 0 | 0 | 0 |
| 9 | 1 | 0 | 0 | 0 | 0 | 1 | 18 | 2 | 0 | 0 | 0 | 1 | 0 |
| 9 | 2 | 1 | 0 | 1 | 0 | 0 | 18 | 3 | 1 | 0 | 1 | 0 | 1 |
| 10 | 1 | 0 | 1 | 0 | 0 | 0 | 19 | 1 | 1 | 0 | 0 | 0 | 0 |
| 10 | 2 | 1 | 0 | 0 | 0 | 1 | 20 | 1 | 0 | 1 | 0 | 0 | 0 |
| 11 | 1 | 0 | 0 | 0 | 0 | 1 | 21 | 1 | 0 | 0 | 1 | 0 | 0 |
| 11 | 2 | 0 | 1 | 1 | 1 | 0 | 22 | 1 | 0 | 0 | 0 | 1 | 0 |
| 12 | 1 | 0 | 0 | 1 | 0 | 0 | 23 | 1 | 0 | 0 | 0 | 0 | 1 |
| 12 | 2 | 1 | 0 | 0 | 1 | 1 |  |  |  |  |  |  |  |

functions for the generating models were the DINA model, DINO model, and *A*-CDM, representing the conjunctive, disjunctive, and additive condensation rules. Note that all categories had the same condensation rule for data generation in each condition. Item quality had three levels: $g = 0.1$, 0.2, or 0.3 for all categories of all items, representing high, moderate, and low quality, where $g = s(\boldsymbol{\alpha}^*_{ljh} = \mathbf{0}) = 1 - s(\boldsymbol{\alpha}^*_{ljh} = \mathbf{1})$ for category $h$ of item $j$. When the *A*-CDM was used as the processing function for data generation, each required attribute was assumed to contribute equally to the processing function. The $Q_c$-matrix is given in Table 1, where each attribute was measured 13 times. The $Q_c$-matrix consists of 6 two-attribute response categories and 6 three-attribute response categories distributed uniformly at Categories 1, 2, and 3. Attribute patterns were generated from the uniform distribution. Like Chalmers, Pek, and Liu (2017), if any information matrices were not positive definite, responses were regenerated. Finally, under each condition, 1,000 valid data sets were analyzed. The GDINA package

(W. Ma & de la Torre, 2017) was used for data simulation and model estimation. The code for model comparison and information matrix calculation was written in the R programming environment (R Core Team, 2017).

## Results

### *Type I Error*

Type I error occurs when a hypothesis test concludes that the G-DINA processing function is statistically better than the generating processing function. For each of the multiattribute response categories, the (observed) Type I error rate is the percentage of times that the hypothesis test makes the Type I error out of the 1,000 replications under a specific significance level. The Type I error rates were averaged across categories with the same $K_{jh}^{*}$ and the level of the response category. The Type I error rates may not be equal to the significance level exactly due to the sampling errors, even when the tests conform well to the nominal level $p$. However, the Type I error rates are expected to have a 95% chance of falling within $p \pm 1.96\sqrt{p(1-p)/n}$ under a nominal level of $p$ with $n$ replications. Researchers are typically interested in .05 $\alpha$ level and because of 1,000 replications under each condition, the observed Type I error rates have a 95% chance of falling within the [.036, .064] interval. Figures 1 through 3 give the Type I error rates of the Wald test using various covariance matrices, two-step LR, and LR tests when $N = 1{,}000$. Figures for Type I error rates when $N = 2{,}000$ and 4,000 are provided as Online Supplemental Materials in the online version of the journal. In these figures, the observed Type I error rates are shown as black bars if they are within this interval and gray bars if not. The line indicating .05 nominal level is also displayed for distinguishing under- and overestimations.

As shown in Figure 1 where $N = 1{,}000$ and items were of high quality, different procedures had varied performance under different conditions. When $K_{jh}^{*} = 2$, all procedures yielded well-calibrated Type I error rates except the Wald test using OPG information matrices, which tended to produce Type I error rates below the nominal level. When $K_{jh}^{*} = 3$, no one procedure is uniformly superior: The Wald test using OPG information matrices still yielded Type I error rates below the nominal level, whereas the Wald test using Louis and sandwich-type information matrices produced Type I error rates above the nominal level under some conditions; both LR test and two-step LR test performed well under most conditions, but yielded inflated Type I error rates under a few conditions. The inflation of the Type I error rates caused by the LR and two-step LR tests were less severe than that caused by the Wald tests using Louis and sandwich information matrices.

From Figure 2 where $N = 1{,}000$ and items were of moderate quality, several findings can be observed. First, the Wald test using OPG information matrices still tended to be conservative under many conditions. Second, when $K_{jh}^{*} = 2$,

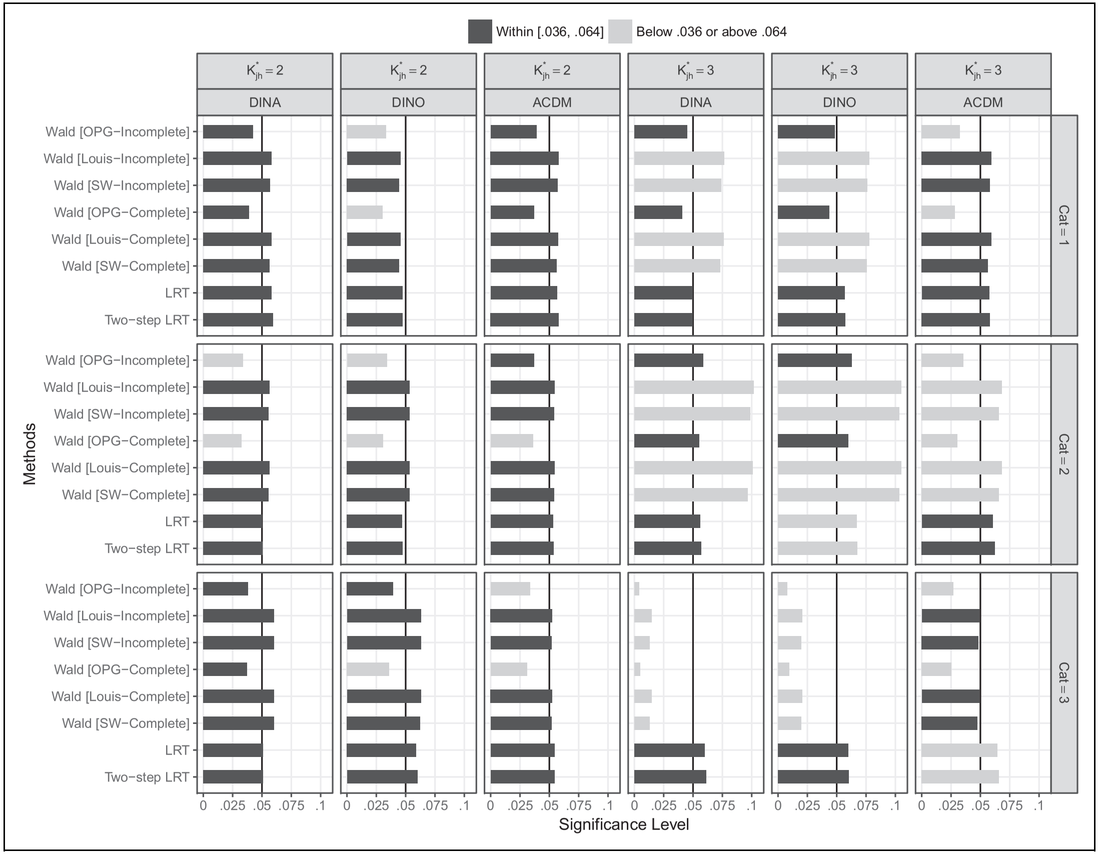

FIGURE 1. *Type I error when* N = *1,000 and items were of high quality.*

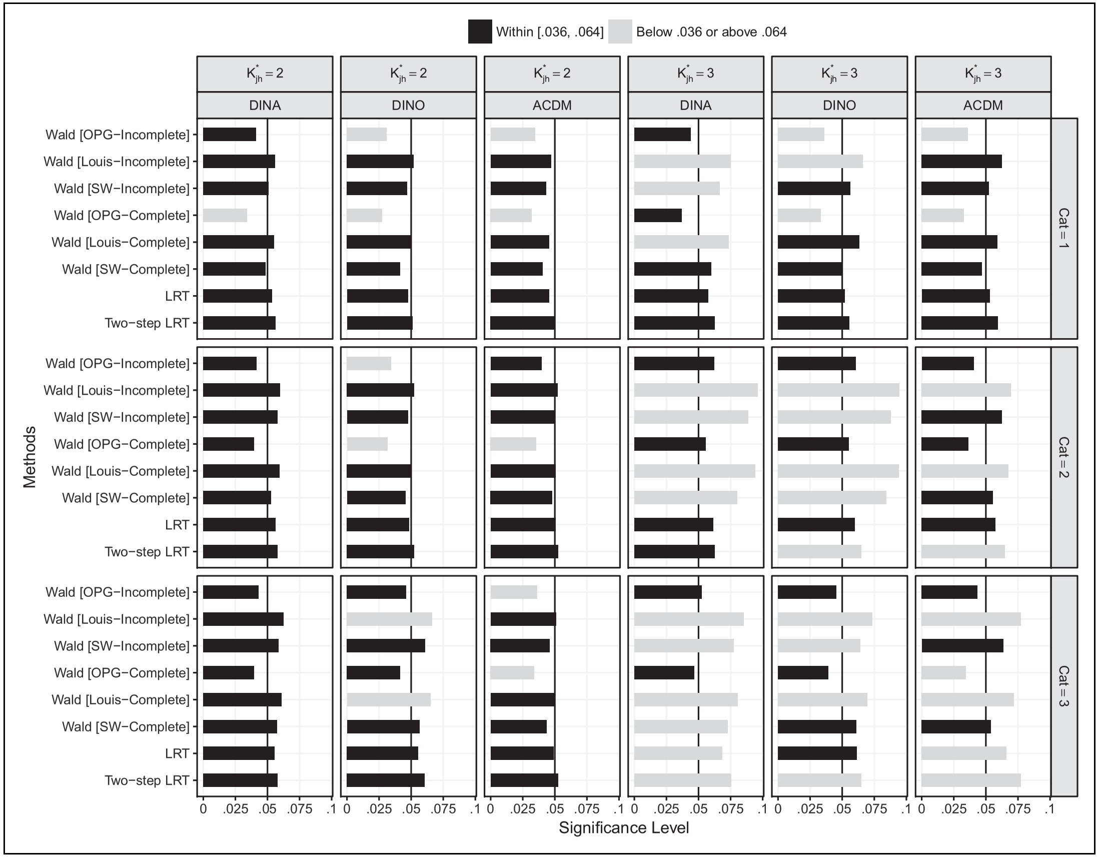

FIGURE 2. *Type I error when* N = *1,000 and items were of moderate quality.*

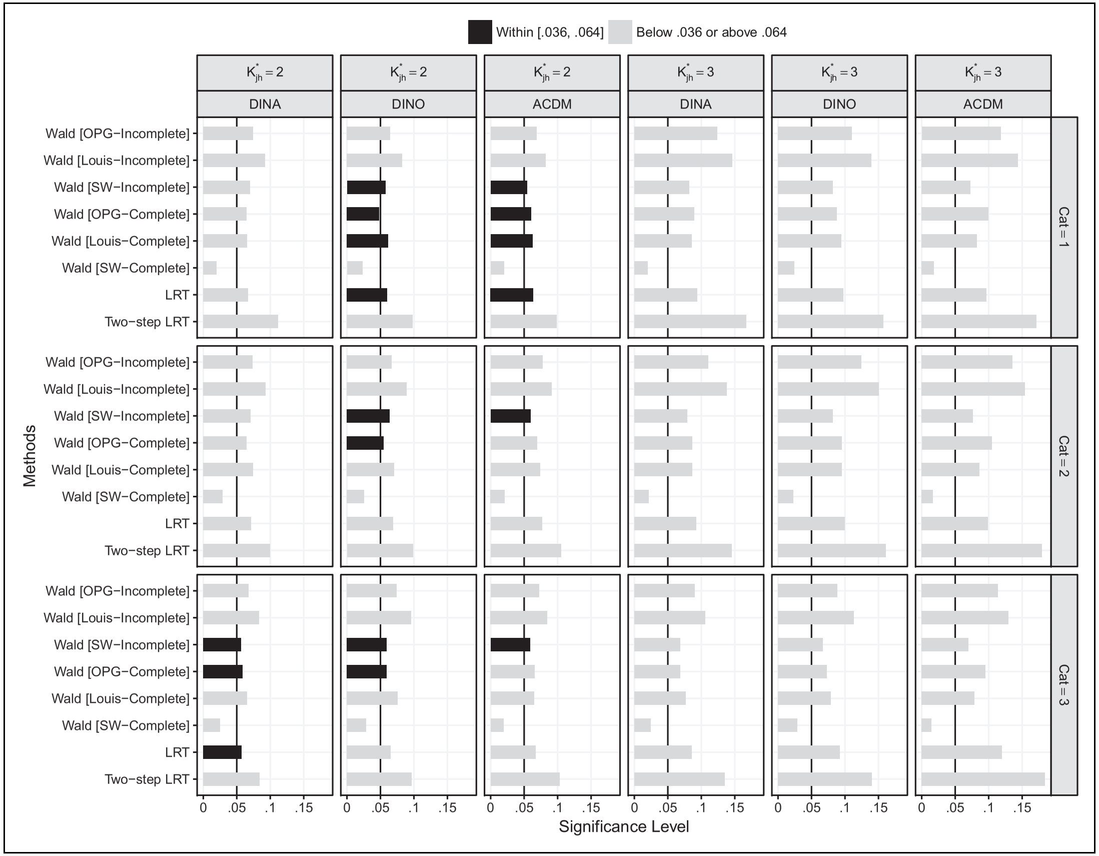

FIGURE 3. *Type I error when* N = *1,000 and items were of low quality.*

other procedures yielded well-calibrated Type I error rates except the Wald test using Louis information matrix, which was slightly liberal for DINO processing function at Category Level 3. Third, when $K_{jh}^{*} = 3$, LR and two-step LR tests, as well as the Wald test using complete sandwich-type information matrix, yielded well-calibrated Type I error rates for Category Level 1; the LR test and the Wald test using OPG information matrices yielded well-calibrated Type I error rates for Category Level 2; and only the Wald test using incomplete OPG information yielded well-calibrated Type I error rates for Category Level 3.

When $N = 1{,}000$ and items were of low quality, as displayed in Figure 3, Type I error rates from these procedures were not well calibrated under most conditions. Unlike other procedures, the Wald test using complete sandwich-type information matrix was conservative.

When $N = 2{,}000$ and items were of moderate or high quality, as displayed in Online Supplemental Figures in the online version of the journal, LR test and two-step LR test outperformed other procedures in general due to well-calibrated Type I error rates under all but one condition. When $N = 2{,}000$ but items were of low quality, two-step LR test yielded more inflated Type I error than most of other procedures, and the Wald test using complete sandwich information matrix produced well-calibrated or slightly underestimated Type I error rates. When $N = 4{,}000$, all procedures were more likely to produce well-calibrated Type I error. In general, LR test seems to be the most reliable in that it produced only one underestimated and one overestimated Type I error of 54 conditions.

### *Empirical Power Rates*

Statistical power indicates the performance of a hypothesis test in rejecting the null hypothesis when it is not correct. To compare statistical power rates, all hypothesis tests should have the same observed Type I error rate. However, this is not the case as shown in the previous section. As a result, the empirical power rates calculated from the empirical distributions under the null hypothesis were reported instead. Specifically, when the generating model was fitted to the data, the fifth percentile of the $p$ values for each hypothesis test was calculated and used as the empirical cutoff for each condition. The empirical power rate, which was calculated for each hypothesis test under each condition, is defined as the percentage of $p$ values that were less than the empirical cutoff under the same condition. Note that if the Type I error rate matches the nominal level, the empirical power rate is the same as the theoretical power. Like the Type I error rate, the empirical power rates were also averaged across categories with the same $K_{jh}^{*}$ and the level of the response category. As in de la Torre and Lee (2013), a test power of 0.80 or higher is considered adequate and of 0.90 or higher is considered excellent. Tables 2 through 4 give the empirical power rates of the Wald and LR tests for the DINA processing functions when the generating processing function was the A-CDM across sample sizes, item qualities, response

TABLE 2.
*Empirical Power for the DINA Processing Function: ACDM–Generated Data*

| | | High Quality | | | Moderate Quality | | | Low Quality | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K_{jh}^*$ | Method | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 |
| 2 | Wald (OPG—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | .996 | .898 | .715 | .442 | .239 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .904 | .720 | .454 | .252 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .908 | .717 | .453 | .258 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 0.999 | 1.000 | .997 | .899 | .711 | .439 | .243 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 1.000 | 1.000 | .996 | .904 | .713 | .448 | .257 |
| | Wald (SW—complete) | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .908 | .638 | .418 | .238 |
| | LR test | 1.000 | 1.000 | 1.000 | 1.000 | .996 | .890 | .698 | .427 | .226 |
| | Two-step LR test | 1.000 | 1.000 | 1.000 | 1.000 | .996 | .890 | .725 | .444 | .231 |
| 3 | Wald (OPG—Incomplete) | 1.000 | 1.000 | 0.913 | 1.000 | .992 | .738 | .599 | .338 | .174 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 0.923 | 1.000 | .991 | .752 | .620 | .351 | .184 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 0.923 | 1.000 | .992 | .759 | .649 | .365 | .192 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 0.905 | 1.000 | .990 | .734 | .600 | .325 | .170 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 0.923 | 1.000 | .992 | .753 | .652 | .338 | .193 |
| | Wald (SW—complete) | 1.000 | 1.000 | 0.924 | 1.000 | .992 | .756 | .648 | .370 | .208 |
| | LR test | 1.000 | 1.000 | 0.940 | 1.000 | .988 | .698 | .617 | .305 | .160 |
| | Two-step LR test | 1.000 | 1.000 | 0.940 | 1.000 | .987 | .692 | .624 | .318 | .163 |

*Note*. $N = 1{,}000$. DINA = deterministic inputs, noisy “and”; ACDM = additive cognitive diagnosis model; OPG = outer product of the gradient; Louis = Louis’s estimator; SW = sandwich-type estimator; Cat = category; LR = likelihood ratio.

TABLE 3.
*Empirical Power for the DINA Processing Function: ACDM–Generated Data*

| | | High Quality | | | Moderate Quality | | | Low Quality | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K_{jh}^*$ | Method | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 |
| 2 | Wald (OPG—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .969 | .785 | .463 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .969 | .787 | .461 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .968 | .790 | .460 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .969 | .782 | .460 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .968 | .784 | .461 |
| | Wald (SW—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .998 | .960 | .784 | .461 |
| | LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .962 | .778 | .444 |
| | Two-step LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .997 | .970 | .799 | .465 |
| 3 | Wald (OPG—incomplete) | 1.000 | 1.000 | 0.998 | 1.000 | 1.000 | .970 | .958 | .664 | .360 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | .970 | .959 | .676 | .371 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | .970 | .961 | .682 | .387 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 0.998 | 1.000 | 1.000 | .970 | .958 | .666 | .360 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | .970 | .962 | .679 | .372 |
| | Wald (SW—complete) | 1.000 | 1.000 | 0.999 | 1.000 | 1.000 | .970 | .963 | .690 | .397 |
| | LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .967 | .954 | .654 | .338 |
| | Two-step LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .967 | .965 | .684 | .366 |

*Note*. $N = 2{,}000$. DINA = deterministic inputs, noisy "and"; ACDM = additive cognitive diagnosis model; OPG = outer product of the gradient; Louis = Louis's estimator; SW = sandwich-type estimator; Cat = category; LR = likelihood ratio.

TABLE 4.
*Empirical Power for the DINA Processing Function: ACDM–Generated Data*

| | | High Quality | | | Moderate Quality | | | Low Quality | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $K_{jh}^*$ | Method | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 | Cat = 1 | Cat = 2 | Cat = 3 |
| 2 | Wald (OPG—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .808 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .810 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .809 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .974 | .807 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .807 |
| | Wald (SW—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .974 | .803 |
| | LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .787 |
| | Two-step LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .978 | .820 |
| 3 | Wald (OPG—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .976 | .726 |
| | Wald (Louis—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .976 | .727 |
| | Wald (SW—incomplete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .976 | .736 |
| | Wald (OPG—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .974 | .725 |
| | Wald (Louis—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .976 | .730 |
| | Wald (SW—complete) | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .975 | .741 |
| | LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .970 | 711 |
| | Two-step LR test | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | .978 | .750 |

*Note*. $N = 4{,}000$. DINA = deterministic inputs, noisy “and”; ACDM = additive cognitive diagnosis model; OPG = outer product of the gradient; Louis = Louis’s estimator; SW = sandwich-type estimator; Cat = category; LR = likelihood ratio.

category levels, and $K_{jh}^*$. The empirical power rates under other conditions can be found in Online Supplemental Materials in the online version of the journal due to the space limits.

From Tables 2 through 4, the empirical power rates of the Wald and LR tests increased as the sample size increased, $K_{jh}^*$ decreased, items quality improved, or the category level decreased. The power rates for all tests were excellent when items were of high quality, with a minimum value of 0.905 occurring when the category level was 3, $N = 1{,}000$, and $K_{jh}^* = 3$. When items were moderate quality, the power rates were higher than 0.967 when $N = 2{,}000$ or higher, but can be as low as 0.692 when $N = 1{,}000$. When items were of low quality, the power rates can be very low especially when the category level was high, and the sample size was small. For example, the power rate for the LR test in distinguishing the DINA processing function from the A-CDM was merely 0.16 when the category level was 3, $N = 1{,}000$, and $K_{jh}^* = 3$. However, increasing the sample size could improve the power substantially. For example, the empirical power rate for the Wald test using the complete OPG information was improved from 0.17 to 0.725 when $N$ increased from 1,000 to 4,000, given that the category level was 3, item quality was low and $K_{jh}^* = 3$.

Similar patterns can be observed under other conditions. The power rates were high under the favorable conditions (i.e., higher item quality, larger sample size, lower level of category, and smaller $K_{jh}^*$) but dropped considerably under some unfavorable conditions. In addition, when the processing function was the DINA model, the power rates for the A-CDM were lower than those for the DINO model under all conditions; and when the processing function was the DINO model, the power rates for the A-CDM were lower than those for the DINA model under all conditions. These results imply that distinguishing the conjunctive and disjunctive models is easier than distinguishing them from the additive model.

When we consider all simulated conditions, no one method outperformed others consistently in terms of the empirical power. However, when the generating processing function is the DINO model, the Wald test regardless of the information matrices produced much lower power rates than the LR and two-step LR tests under high item quality, small sample size, $K_{jh}^* = 3$, and the category level was 3. For example, the power rates of the Wald test using different information matrices for the A-CDM ranged from 0.662 to 0.700 when $N = 1{,}000$, $K_{jh}^* = 3$, and the category level was 3. In contrast, under the same condition, the power rates of the LR and two-step LR tests were 0.925.

## Real Data Analysis

Responses of 1,328 students from the United States to 17 items from the Block 4 of the TIMSS 2007 eighth-grade mathematics assessment were analyzed in this

TABLE 5.
$Q_c$-*Matrix for the Trends in International Mathematics and Science Study 2007 Data*

| Item No. | TIMSS Item ID | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ |
|---|---|---|---|---|---|---|---|---|
| 1 | M042001 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | M042022 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | M042082 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | M042088 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 5 | M042304A | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6-1 | M042304B-1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 6-2 | M042304B-2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 7 | M042304C | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 8-1 | M042304D-1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8-2 | M042304D-2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | M042267 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 10 | M042239 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 11 | M042238 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 12 | M042279 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 13 | M042036 | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| 14 | M042130 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| 15 | M042303A | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 16-1 | M042303B-1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 16-2 | M042303B-2 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| 17 | M042222 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |

study. The attributes measured by these items were identified by L. Ma (2014), who considered both cognitive process attributes and content attributes and built attributes at two levels. However, for illustration purposes, only seven second-level content attributes were considered in this study, namely, ($\alpha_1$) whole numbers and integers; ($\alpha_2$) fractions, decimals, ratio proportion, and percent; ($\alpha_3$) algebraic expressions and equations/formulas functions; ($\alpha_4$) geometric shapes; ($\alpha_5$) geometric measurement and location movement; ($\alpha_6$) data organization and representation; and ($\alpha_7$) data interpretation and chance. L. Ma (2014) also developed the Q-matrix for these items using multiple regression and the least squares distance method, and the $Q_c$-matrix, given in Table 5, was created based on L. Ma's work by assuming that for each polytomously scored item, all required attributes are measured by each step of the item. The sequential G-DINA model was fitted to the data. The Wald test using the incomplete and complete OPG, Louis, and sandwich information matrices; the LR; and the two-step LR tests were conducted to examine whether the saturated G-DINA model can be replaced by the DINA model, DINO model, and A-CDM. However, the complete and incomplete Louis's information matrices were not

TABLE 6.
p *Values of the Wald and LR Tests for the ACDM Processing Function*

| | Wald | Wald | Two-Step | |
|---|---|---|---|---|
| Item No. | (OPG—Incomplete) | (OPG—Complete) | LR Test | LR Test |
| 3 | | | | |
| 6-1 | | | | |
| 6-2 | | 0.083 | | |
| 7 | 0.282 | 0.481 | 0.140 | 0.684 |
| 9 | 0.501 | 0.543 | 0.440 | 0.543 |
| 10 | | | | |
| 11 | | | | |
| 13 | 0.635 | 0.723 | 0.304 | 0.570 |
| 14 | | 0.216 | | |
| 15 | | 0.081 | | |
| 16-1 | 0.137 | 0.136 | | |
| 16-2 | 0.092 | 0.343 | | 0.275 |
| 17 | 1.000 | 1.000 | 1.000 | 1.000 |

*Note. p* values less than .05 were omitted. ACDM = additive cognitive diagnosis model; OPG = outer product of the gradient; LR = likelihood ratio.

positive definite, and therefore both Louis's and sandwich covariance matrices were not calculable.

For 13 multi-attribute response categories from 9 dichotomous items and 2 polytomous items, the DINA model was never selected by any of the hypothesis tests. The DINO model was selected only once for Item 17 by the Wald test using the complete information matrix with a *p* value of .51; whereas for this item, A-CDM was believed appropriate by all hypothesis tests. Because the DINA and DINO models were not selected for any other category, only the *p* values for A-CDM were given in Table 6. The results from all hypothesis tests were consistent for 8-item response categories. Specifically, the G-DINA model was deemed appropriate for Items 3, 6-1, 10, and 11, whereas the sequential A-CDM was considered as good as the sequential G-DINA model for Items 7, 9, 13, and 17. For other items or categories, different procedures selected different condensation rules. Based on the Wald test using the complete information matrix, the A-CDM was appropriate for 9-item response categories; whereas based on two-step LR test, A-CDM may only be used for four response categories.

According to the models suggested by each hypothesis test, the data were recalibrated, and the AIC and BIC were calculated for each fitted model. The LR test was also implemented at test level to evaluate whether the suggested models fitted as well as the saturated sequential G-DINA model. Based on the test-level

TABLE 7.
*AIC and BIC for Models Selected by the Wald and LR Tests*

| | Wald<br>(OPG—Incomplete) | Wald<br>(OPG—Complete) | Two-Step<br>LR Test | LR Test |
|---|---|---|---|---|
| AIC | **27,419.739** | 27,438.274 | 27,431.159 | 27,420.636 |
| BIC | 28,499.556 | **28,450.602** | 28,625.188 | 28,557.559 |

*Note*. Lowest AIC and BIC were shown in boldface. AIC and BIC for the sequential generalized deterministic inputs, noisy "and" gate model were 27,441.857 and 28,672.226, respectively. OPG = outer product of the gradient; LR = likelihood ratio; AIC = Akaike information criterion; BIC = Bayesian information criterion.

LR test, the models suggested by the Wald test using the complete information matrix were significantly worse than the sequential G-DINA model, $\chi^2 = 80.417, df = 42, p < .001$, whereas the models suggested by other hypothesis tests were all statistically as good as the sequential G-DINA model. From Table 7, if not taking the Wald test using the complete information into consideration, the models based on the Wald test using the incomplete information had the smallest AIC and BIC, followed by those suggested by the LR test.

Table 8 gives the classification consistency among a variety of sequential CDMs, including the sequential G-DINA, DINA, DINO, and A-CDM, as well as the sequential models selected using Wald and LR tests. The attribute profiles were estimated using the expected a posteriori method. The upper triangle shows the classification consistency at attribute level, which is defined as the proportion of individual attributes that were identically classified by two CDMs, whereas the lower triangle shows the classification consistency at attribute vector level, which is defined as the proportion of individuals who were classified into the same latent class by two CDMs. It can be observed that the condensation rules can have a strong influence on individual classifications. For example, the classification consistency at attribute vector level ranged from 0.284 to 0.785 among the sequential G-DINA, DINA, DINO, and A-CDM. The model selected by the two-step LR test had the highest classification consistency (i.e., 0.925) with the sequential G-DINA model, followed by the model selected by the LR test (i.e., 0.893). It should be noted that the classification consistency results show the impact of condensation rules on individual classifications but do not tell us which model is the best in that the true attribute profile for each individual is unknown.

## Discussion

It has been said that no model is true, but some are more useful than others. A psychometric model should be in line with the underlying cognitive processes to provide a good approximation to the reality. The condensation rule is a central

TABLE 8.
*Classification Consistency Among Different CDMs*

| Models | sG-DINA | sDINA | sDINO | sA-CDM | Wald (Incomplete) | Wald (Complete) | Two-Step LR Test | LR Test |
|---|---|---|---|---|---|---|---|---|
| sG-DINA | | .865 | .806 | .960 | .976 | .964 | .988 | .983 |
| sDINA | .416 | | .769 | .868 | .860 | .865 | .863 | .863 |
| sDINO | .337 | .284 | | .813 | .809 | .809 | .808 | .808 |
| sA-CDM | .785 | .390 | .367 | | .968 | .973 | .965 | .965 |
| Wald (OPG—incomplete) | .860 | .374 | .355 | .828 | | .972 | .900 | .986 |
| Wald (OPG—complete) | .783 | .368 | .365 | .846 | .831 | | .809 | .969 |
| Two-step LR test | .925 | .398 | .338 | .809 | .983 | .968 | | .989 |
| LR test | .893 | .385 | .340 | .809 | .921 | .817 | .935 | |

*Note*. Lower triangle shows the classification consistency at attribute vector level and upper triangle shows the classification consistency at attribute level. OPG = outer product of the gradient; sGDINA = sequential generalized deterministic inputs, noisy “and” gate model; sDINA = seqential deterministic inputs, noisy “and” model; sDINO = sequential deterministic inputs, noisy “or” model; sA-CDM = sequential additive cognitive diagnosis model; LR test = likelihood ratio test.

component for many CDMs, and in this study, we examined the Type I error and power of the Wald and LR tests in determining the appropriate condensation rules for each response category of a polytomously scored item. This is achieved by comparing whether the reduced models can be used in place of the G-DINA model without a significant loss in model-data fit.

This study systematically examined the influence of incomplete and complete OPG, Louis's, and sandwich covariance matrices on the Wald test. Results show that the Wald test using complete or incomplete OPG information matrices tended to be conservative especially when sample size was small. This finding is partially consistent with Y. Liu, Xin, Li, Tian, and Liu (2016) who examined the Type I error rates of the Wald test for the DINA model using the complete OPG information matrix in detecting differential item functioning for dichotomous responses and found that the Wald test tended to be conservative under small sample sizes regardless of item quality.

The Wald test based on the Louis's information matrices does not outperform that based on the simpler OPG information matrices. Also, the Louis's information matrices are more likely to be nonpositive definite as noticed by other researchers (e.g., Duan & Fulop, 2011). Despite this issue, the resulting sandwich information matrices could help the Wald test control the inflation of the Type I error when items were of low quality.

Despite not involving an estimated covariance matrix, the LR test also produced inflated Type I error rates under some unfavorable conditions. However, unlike the Wald test, the LR test did not tend to be conservative under high item quality conditions. Although the two-step LR test performed as well as the LR test under most conditions, it tends to yield much more inflated Type I error rates than the LR test when items were of low quality and sample size was small.

In terms of the computation time, the LR test can be very expensive if the data calibration takes time or the number of categories is very large. For the real data analyzed in this study, the Wald test using the OPG information took only about 0.25 s to compare the G-DINA processing function with the DINA, DINO, and A-CDM for all multiattributes categories. In contrast, the LR test and two-step LR test took around 16 min and 12 s, respectively. It should be noted that the code for the LR test was written in R by the author, and faster speeds can be expected by using a program written in a lower level language such as C. Under most conditions, the studied methods have similar power rates, and no single method performed best across all conditions. Despite excellent power rates under favorable conditions, their power can drop substantially under unfavorable conditions.

Based on the closed-form solution for item parameter estimation, the processing function for category *h* is the ratio of the expected number of examinees, given a particular attribute pattern obtaining a score of *h* or higher to the expected number of examinees and given the attribute pattern obtaining a score of $h - 1$ or higher. As a result, the number of examinees who get at least a score of $h - 1$ can be viewed as the "effective" sample size for category $h$ and thus for a higher category, the "effective" sample size is smaller, yielding a poorer power. A related finding is given by W. Ma (2018), where the sequential G-DINA model is a special case of the diagnostic tree model and the lower categories tend to have better parameter recovery due to the larger effective samples.

A set of data from the TIMSS assessment was retrofitted to illustrate the use of the Wald and LR tests in practice. It is shown that, for the current data, models with different condensation rules produced substantial different person classifications, and thus, it is important to choose CDMs that can approximate the underlying cognitive processes well using, for example, the Wald and LR tests. However, the results need to be interpreted with cautions because of many challenges associated with retrofitting (R. Liu, Huggins-Manley, & Bulut, 2017). For example, many attributes are relatively coarser-grained but still assumed to be binary latent variables for CDM analyses. This may produce unstable or inaccurate parameter estimates. It would be interesting to explore how attributes should be defined to balance the grain size and the number of attributes in the context of large-scale assessments as in Skaggs, Wilkins, and Hein (2016).

It is worth emphasizing that retrofitting tends to be suboptimal. This study intends to provide a set of tools that can be used along with the sequential G-DINA model, so that researchers can develop the CDAs under this framework. To achieve this goal, future research along this line is needed. For example, the test length and the number of attributes were fixed and the Q-matrix was assumed known. Guo, Ma, and de la Torre (2017) found that with misspecified Q-matrix, the standard error of item parameters estimated using the OPG approximation can be problematic, so future research may examine their influence on the performance of the Wald test. In addition, apart from the OPG, Louis's, and sandwich information matrices investigated in this study, future research may explore the performance of the Wald test using covariance matrix calculated in other ways, such as the Oakes's method (Chalmers, 2018) and the numerical differential methods (Jamshidian & Jennrich, 2000).

## Appendix A

### Score Functions for the OPG Information

To calculate the information matrix using OPG method, Equation 13 involves the observed data score function with respect to the processing function $s(\boldsymbol{\alpha}_{ljh}^{*})$, which can be written as

$$
\begin{aligned}
\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})} &= \frac{1}{f(\boldsymbol{Y}_i;\boldsymbol{\psi})}\frac{\partial f(\boldsymbol{Y}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})} \\
&= \frac{1}{f(\boldsymbol{Y}_i;\boldsymbol{\psi})}\frac{\partial}{\partial s(\boldsymbol{\alpha}^*_{ljh})}\sum_c \pi_c f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi}) \\
&= \sum_c \frac{\pi_c}{f(\boldsymbol{Y}_i;\boldsymbol{\psi})}\frac{\partial f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})} \\
&= \sum_c \frac{\pi_c f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi})}{f(\boldsymbol{Y}_i;\boldsymbol{\psi})}\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})} \\
&= \sum_c f(\boldsymbol{\alpha}_c;\boldsymbol{Y}_i,\boldsymbol{\psi})\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})}.
\end{aligned} \tag{A1}
$$

For the sequential G-DINA model, we have

$$
\begin{aligned}
\log f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi}) &= \sum_{j=1}^{J}\sum_{h=0}^{H_j} I(Y_{ij}=h)\log P(Y_{ij}=h;\boldsymbol{\alpha}_c,\boldsymbol{\psi}) \\
&= \sum_{j=1}^{J}\sum_{h=1}^{H_j}\{I(Y_{ij}=h-1)\log[1-s_{jh}(\boldsymbol{\alpha}_c)]+I(Y_{ij}\geq h)\log[s_{jh}(\boldsymbol{\alpha}_c)]\}.
\end{aligned} \tag{A2}
$$

Therefore, Equation A1 can be written as

$$
\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})} = f(\boldsymbol{\alpha}^*_{ljh};\boldsymbol{Y}_i)\left[\frac{I(Y_{ij}\geq h)}{s(\boldsymbol{\alpha}^*_{ljh})}-\frac{I(Y_{ij}=h-1)}{1-s(\boldsymbol{\alpha}^*_{ljh})}\right]. \tag{A3}
$$

In addition, for complete information matrix using the OPG method, we need the observed data score function for the latent class proportion parameters $\boldsymbol{\pi}$, which, as shown in Philipp et al. (2018), has elements

$$
\frac{\partial \log f(\boldsymbol{Y}_i;\boldsymbol{\psi})}{\partial \pi_c} = \frac{f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_c,\boldsymbol{\psi})-f(\boldsymbol{Y}_i;\boldsymbol{\alpha}_1,\boldsymbol{\psi})}{f(\boldsymbol{Y}_i;\boldsymbol{\psi})}. \tag{A4}
$$

## Appendix B

### Derivations for Louis's Information Matrix

With independent observations, it follows from Equation (3.2′) of Louis (1982) that

$$
\begin{aligned}
\mathcal{I}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}) = &\sum_{i=1}^{N}\int \boldsymbol{B}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i - \sum_{i=1}^{N}\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})\boldsymbol{S}'(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i \\
&-2\sum_{i=1}^{N-1}\sum_{i'=i+1}^{N}\left[\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i\right]\left[\int \boldsymbol{S}'(\boldsymbol{X}_{i'};\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_{i'};\boldsymbol{Y}_{i'})d\boldsymbol{a}_{i'}\right],
\end{aligned} \tag{B1}
$$

where $\boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})$ and $\boldsymbol{B}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})$ are score function and negative hessian matrix, respectively, as defined in Equation 15. Since

$$\boldsymbol{S}(\boldsymbol{Y};\widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{N}\boldsymbol{S}(\boldsymbol{Y}_i;\widehat{\boldsymbol{\psi}}) = \sum_{i=1}^{N}\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i = 0, \tag{B2}$$

we have

$$\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i = -\sum_{i'=1,i'\neq i}^{N}\int \boldsymbol{S}(\boldsymbol{X}_{i'};\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_{i'};\boldsymbol{Y}_{i'})d\boldsymbol{a}_{i'}. \tag{B3}$$

After a few algebraic manipulations, Equation B1 can be written as

$$\begin{aligned}\mathcal{I}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}) &= \sum_{i=1}^{N}\int \boldsymbol{B}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i - \sum_{i=1}^{N}\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})\boldsymbol{S}'(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i\\ &+\sum_{i=1}^{N}\left[\int \boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i\right]\left[\int \boldsymbol{S}'(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)d\boldsymbol{a}_i\right].\end{aligned} \tag{B4}$$

Since $a_{ik}$ is assumed to be a binary variable, Equation B4 can be written as

$$\begin{aligned}\mathcal{I}_{\text{Louis}}(\widehat{\boldsymbol{\psi}}) &= \sum_{i=1}^{N}\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{B}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)\\ &-\sum_{i=1}^{N}\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})\boldsymbol{S}'(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)\\ &+\sum_{i=1}^{N}\left[\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{S}(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)\right]\left[\sum_{a_{i1}=0}^{1}\cdots\sum_{a_{iK}=0}^{1}\boldsymbol{S}'(\boldsymbol{X}_i;\widehat{\boldsymbol{\psi}})f(\boldsymbol{a}_i;\boldsymbol{Y}_i)\right].\end{aligned} \tag{B5}$$

For the sequential G-DINA model, we have

$$\begin{aligned}\log f(\boldsymbol{X}_i;\boldsymbol{\psi}) &= \log f(\boldsymbol{Y}_i,\boldsymbol{a}_i;\boldsymbol{\psi}) = \log f(\boldsymbol{Y}_i;\boldsymbol{a}_i,\boldsymbol{\psi}) + \log f(\boldsymbol{a}_i)\\ &= \sum_{j=1}^{J}\sum_{h=0}^{H_j} I(Y_{ij}=h)\log P(Y_{ij}=h;\boldsymbol{a}_i,\boldsymbol{\psi}) + \log f(\boldsymbol{a}_i)\\ &= \sum_{j=1}^{J}\sum_{h=1}^{H_j}\{I(Y_{ij}=h-1)\log[1-s_{jh}(\boldsymbol{a}_i)] + I(Y_{ij}\geq h)\log[s_{jh}(\boldsymbol{a}_i)]\} + \log f(\boldsymbol{a}_i).\end{aligned} \tag{B6}$$

The score function $\boldsymbol{S}(\boldsymbol{X}_i;\boldsymbol{\psi})$ has elements of

$$\frac{\partial \log f(\boldsymbol{X}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}_{ljh}^{*})} = \begin{cases}\dfrac{I(Y_{ij}\geq h) - s(\boldsymbol{\alpha}_{ljh}^{*})}{s(\boldsymbol{\alpha}_{ljh}^{*})[1-s(\boldsymbol{\alpha}_{ljh}^{*})]} & \text{if } \boldsymbol{a}_i \in C_{ljh}\\ 0 & \text{otherwise}\end{cases}, \tag{B7}$$

$$\frac{\partial \log f(\boldsymbol{X}_i;\boldsymbol{\psi})}{\partial \pi_c} = \begin{cases} -\dfrac{1}{\pi_1} & \text{if } \boldsymbol{a}_i = \boldsymbol{\alpha}_1 \\ \dfrac{1}{\pi_c} & \text{if } \boldsymbol{a}_i = \boldsymbol{\alpha}_c, \text{ and } c \neq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{B8}$$

and $\boldsymbol{B}(\boldsymbol{X}_i;\boldsymbol{\psi})$ has elements of

$$-\frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})\partial s(\boldsymbol{\alpha}^*_{l'jh})} = \begin{cases} \dfrac{I(Y_{ij} \geq h)}{[s(\boldsymbol{\alpha}^*_{ljh})]^2} + \dfrac{I(Y_{ij} = h-1)}{[1 - s(\boldsymbol{\alpha}^*_{ljh})]^2} & \text{if } \boldsymbol{a}_i \in C_{ljh}, \text{and } l = l' \\ 0 & \text{otherwise} \end{cases}, \tag{B9}$$

$$-\frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\psi})}{\partial \pi_c \partial \pi_{c'}} = \begin{cases} \dfrac{1}{\pi_1^2} & \text{if } \boldsymbol{a}_i = \boldsymbol{\alpha}_1 \\ \dfrac{1}{\pi_c^2} & \text{if } \boldsymbol{a}_i = \boldsymbol{\alpha}_c, c = c' \neq 1 \\ 0 & \text{otherwise} \end{cases}, \tag{B10}$$

$$-\frac{\partial^2 \log f(\boldsymbol{X}_i;\boldsymbol{\psi})}{\partial s(\boldsymbol{\alpha}^*_{ljh})\partial \pi_c} = 0. \tag{B11}$$

## Authors' Note

This article is based on the first author's PhD dissertation at Rutgers, The State University of New Jersey, under the supervision of the second author.

## Declaration of Conflicting Interests

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

## Funding

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

## ORCID iD

W. Ma http://orcid.org/0000-0002-6763-0707

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Chalmers, R. P. (2018). Numerical approximation of the observed information matrix with Oakes' identity. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12127

Chalmers, R. P., Pek, J., & Liu, Y. (2017). Profile-likelihood confidence intervals in item response theory models. *Multivariate Behavioral Research*, *52*, 533–550.

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.

de la Torre, J. (2008). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

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.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). A review of cognitively diagnostic assessment and a summary of psychometric models. In R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Psychometrics* (Vol. 26, pp. 979–1030). Amsterdam, the Netherlands: Elsevier.

Duan, J. C., & Fulop, A. (2011). A stable estimator of the information matrix under EM for dependent data. *Statistics and Computing*, *21*, 83–91. doi:10.1007/s11222-009-9149-4

Guo, W., Ma, W., & de la Torre, J. (2017). *Standard error estimation using bootstrap approaches for cognitive diagnosis models*. Paper presented at the Annual Meeting of the National Council of Measurement in Education, San Antonio, TX.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign, Urbana–Champaign.

Hemker, B. T., van der Ark, L. A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*, 487–506.

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*, 191–210.

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98–125.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In L. Le Cam & J. Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 221–233). Berkeley, CA: University of California Press.

Jamshidian, M., & Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*, 257–270.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*, 59–81.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Liu, R., Huggins-Manley, A. C., & Bulut, O. (2017). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement*. doi:10.1177/0013164416685599

Liu, Y., Tian, W., & Xin, T. (2016). An application of M2 statistic to evaluate the fit of cognitive diagnostic models. *Journal of Educational and Behavioral Statistics*, *41*, 3–26.

Liu, Y., & Xin, T. (2017). *dcminfo: Information matrix for diagnostic classification models* (R package version 0.1.7) [Computer software]. Retrieved from https://CRAN.R-project.org/package=dcminfo

Liu, Y., Xin, T., Andersson, B., & Tian, W. (2018). Information matrix estimation procedures for cognitive diagnostic models. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12134

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.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *44*, 226–233.

Ma, L. (2014). *Validation of the item-attribute matrix in TIMSS: Mathematics using multiple regression and the LSDM* (Unpublished doctoral dissertation). University of Denver, Denver, CO.

Ma, W. (2018). A diagnostic tree model for polytomous responses with multiple strategies. *British Journal of Mathematical and Statistical Psychology*. doi:10.1111/bmsp.12137

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. (2017). *GDINA: The generalized DINA model framework* [R package version 1.4.2 computer software]. Retrieved from https://CRAN.R-project.org/package=GDINA

Ma, W., Iaconangelo, C., & de la Torre, J. (2016). Model similarity, model selection, and attribute classification. *Applied Psychological Measurement*, *40*, 200–217.

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

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Paek, I., & Cai, L. (2014). A comparison of item parameter standard error estimation procedures for unidimensional and multidimensional item response theory modeling. *Educational and Psychological Measurement*, *74*, 58–76.

Philipp, M., Strobl, C., de la Torre, J., & Zeileis, A. (2018). On the estimation of standard errors in cognitive diagnosis models. *Journal of Educational and Behavioral Statistics*, *43*, 88–115.

R Core Team. (2017). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rojas, G., de la Torre, J., & Olea, J. (2012). *Choosing between general and specific cognitive diagnosis models when the sample size is small.* Paper presented at the

Annual Meeting of the National Council of Measurement in Education, Vancouver, Canada.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sinharay, S., & Almond, R. G. (2007). Assessing fit of cognitive diagnostic models: A case study. *Educational and Psychological Measurement*, *67*, 239–257.

Skaggs, G., Wilkins, J. L., & Hein, S. F. (2016). Grain size and parameter recovery with TIMSS and the general diagnostic MODEL. *International Journal of Testing*, *16*, 310–330.

Sorrel, M. A., de la Torre, J., Abad, F. J., & Olea, J. (2017). Two-step likelihood ratio test for model comparison in cognitive diagnosis models. *Methodology*, *13*, 39–47.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 583–639.

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

Tutz, G. (2016). Sequential models for ordered responses. In W. J. van der Linden (Ed.), *Handbook of item response theory (Vol. 1: Models)* (pp. 139–150). Boca Raton, FL: CRC Press.

Verhelst, N. D., Glas, C. A. W., & de Vries, H. H. (1997). A step model to analyze partial credit. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 123–138). New York, NY: Springer-Verlag.

Verhelst, N. D., & Verstralen, H. H. (2008). Some considerations on the partial credit model. *Psicologica*, *29*, 229–254.

von Davier, M. (2008). A general diagnostic model applied to language testing data. *British Journal of Mathematical and Statistical Psychology*, *61*, 287–307.

von Davier, M. (2014). The DINA model as a constrained general diagnostic model: Two variants of a model equivalency. *British Journal of Mathematical and Statistical Psychology*, *67*, 49–71.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*, 1–25.

Xin, T., Liu, Y., Tian, W., & Li, L. (2017, April). *New item-level model selection procedures for diagnostic classification models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Antonio, TX.

Yuan, K. H., Cheng, Y., & Patton, J. (2014). Information matrices and standard errors for MLEs of item parameters in IRT. *Psychometrika*, *79*, 232–254.

## Authors

WENCHAO MA is an assistant professor at the College of Education at The University of Alabama, Box 870231, Tuscaloosa, AL 35487, USA; email: wenchao.ma@ua.edu. His research interests include educational and psychological measurement, cognitive diagnosis modeling, and item response theory.

JIMMY DE LA TORRE is a professor at the Faculty of Education at The University of Hong Kong, Pokfulam Road, Hong Kong, Hong Kong; email: j.delatorre@hku.hk. His research interests include psychometrics, item response models, cognitive diagnosis models, and the use of assessments to inform classroom instruction and learning.

Manuscript received September 23, 2017
First revision received February 4, 2018
Second revision received May 5, 2018
Accepted July 5, 2018