Journal of Educational Measurement
Winter 2013, Vol. 50, No. 4, pp. 355–373

Evaluating the Wald Test for Item-Level Comparison of Saturated and Reduced Models in Cognitive Diagnosis

Jimmy de la Torre
Rutgers, The State University of New Jersey
Young-Sun Lee
Teachers College–Columbia University

This article used the Wald test to evaluate the item-level fit of a saturated cognitive diagnosis model (CDM) relative to the fits of the reduced models it subsumes. A simulation study was carried out to examine the Type I error and power of the Wald test in the context of the G-DINA model. Results show that when the sample size is small and a larger number of attributes are required, the Type I error rate of the Wald test for the DINA and DINO models can be higher than the nominal significance levels, while the Type I error rate of the A-CDM is closer to the nominal significance levels. However, with larger sample sizes, the Type I error rates for the three models are closer to the nominal significance levels. In addition, the Wald test has excellent statistical power to detect when the true underlying model is none of the reduced models examined even for relatively small sample sizes. The performance of the Wald test was also examined with real data. With an increasing number of CDMs from which to choose, this article provides an important contribution toward advancing the use of CDMs in practical educational settings.

In recent years, cognitive diagnostic models (CDMs) have received growing attention and interest among researchers and practitioners in the field of educational testing and measurement because of their potential to improve classroom instruction and learning. CDMs are psychometric models involving multiple discrete latent variables developed to provide specific information about the cognitive skills or attributes required to solve problems in a particular domain. Put differently, CDMs can be used to identify students' specific strengths and weaknesses in a domain of interest. As such, the finer grained information derived using CDMs from appropriately constructed assessments can be more diagnostically useful and relevant in many instructional settings. In contrast, most commonly employed psychometric frameworks, as in classical test theory and item response theory (IRT), provide single overall scores that indicate a student's relative position on an ability continuum. As noted by de la Torre and Karelitz (2009), efforts have been exerted to make unidimensional tests more interpretative and diagnostic. However, without providing supplementary information (e.g., representative items at specific ability levels), these scores in a traditional reporting format would be mainly informative for accurate rank ordering of students and would be of limited diagnostic value.

At present, several specific and general CDMs of various formulations exist in the psychometric literature. Examples of specific CDMs include the deterministic inputs, noisy "and" gate model (NIDA: Haertel, 1989; Junker & Sijtsma, 2001), the deterministic inputs, noisy "or" gate model (NIDO; Templin & Henson, 2006), and the

reduced reparametrized unified model (R-RUM; DiBello, Roussos, & Stout, 2007; Hartz, 2002; Templin, 2006); examples of general CDMs include the generalized DINA model (G-DINA; de la Torre, 2011), the log-linear CDM (Henson, Templin, & Willse, 2009), and the general diagnostic model (GDM; von Davier, 2005). Although general CDMs provide better model–data fit compared to specific CDMs, they are also more complex and thus require larger sample sizes to be estimated accurately. In addition, specific CDMs have more straightforward and meaningful interpretations. Also, the parsimony principle dictates that the simplest model be chosen when a set of statistically indistinguishable models is available. Finally, Rojas, de la Torre, and Olea (2012) showed that, compared to a general CDM, using the correct specific CDMs can lead to higher correct attribute classification rates, particularly when the sample size is small.

With multiple options, the best way to choose the most appropriate model from the available CDMs apart from personal preference is not clear. The overarching goal of this article is to examine the viability of a statistical test—the Wald test—as a more objective means of identifying the most appropriate CDM from a collection of CDMs. Specifically, this article will document the Type I error and power of the Wald test in comparing the fit of a saturated (i.e., general) model against the fits of the reduced models it subsumes at the item level. The properties of the Wald test are investigated in the context of the G-DINA model.

When several viable models are available for fitting the data, selecting the most appropriate model becomes a critical issue. The process of model selection involves checking the model–data fit, which is a multifaceted endeavor that can be examined at the test or item level and at the person level. To a large extent, the process of determining the most appropriate model is a validation process. Thus, as with most procedures for validating theories in scientific investigations, model selection often is conducted by comparing the theory-based predictions and actual observations. In the psychometric context, theory can be narrowly construed as the specific form of the psychometric model used to analyze the data. If some important characteristics of the data can be reproduced (e.g., first and second moments), the model is said to fit the data (e.g., see Hambleton & Han, 2005). As such, the posited underlying process articulated in the psychometric model provides a plausible explanation for the observed data.

CDM fit evaluation can be classified in various ways. As in traditional IRT (see Swaminathan, Hambleton, & Rogers, 2007), one way of classifying CDM fit evaluation is to distinguish whether evaluation is done at the item or test level. Test-level fit evaluation refers to the process of determining whether the posited model fits the data in their entirety; in contrast, item-level fit evaluation refers to the process of examining whether the model fits the individual items. It should be noted that some procedures for test-level fit evaluation are based on aggregated item-level statistics. Another way of classifying CDM fit evaluation is to determine whether fit is evaluated in absolute or relative terms (Chen, de la Torre, & Zhang, 2013). In the former, model–fit evaluation is used to examine whether the model is appropriate for the data; in the latter, model–fit evaluation is used to determine which model to choose from a set of several competing models.

Several of the statistics and procedures used in CDM fit evaluation based on conventional fit statistics (i.e., Akaike's information criterion [AIC], Akaike, 1973; Bayesian information criterion [BIC], Schwarz, 1976; deviance information criterion [DIC], Spiegelhalter, Best, Carlin, & van der Linde, 2002; Bayes factor, Kass & Raftery, 1995) have been used by several researchers (e.g., de la Torre & Douglas, 2004, 2008; DeCarlo, 2011; Henson et al., 2009; Rupp, Templin, & Henson, 2010; Sinharay & Almond, 2007) to evaluate relative fit at the test level. Examples of absolute fit evaluations at the test level include statistics based on the residuals between the estimated and observed first two moments (e.g., de la Torre & Douglas, 2008; Henson et al., 2009; Kunina-Habenicht, Rupp, & Wilhelm, 2012; Rupp et al., 2010; Sinharay & Almond, 2007).

The above statistics and procedures are all carried out to facilitate test-level decisions. Missing are procedures for making decisions at the item level, both in relative and absolute terms. This article focuses on how the relative fit of CDMs can be evaluated at the item level. Specifically, a general model (i.e., G-DINA model) can be tested statistically against the fits of some of the specific CDMs it subsumes using the Wald test. The Wald test was originally proposed by de la Torre (2011) for comparing general and specific models at the item level (i.e., one item at a time) creating the possibility of using multiple CDMs within the same test. However, before the Wald test can be adopted as a method for determining the most appropriate models in applied testing settings, it is important to examine its statistical properties, particularly its Type I error rate and power with respect to three special cases or reduced forms of the G-DINA model (i.e., DINA model, DINO model, and *A*-CDM).

Background

The G-DINA Model

The DINA model is one of, if not the simplest, interpretable CDMs appropriate for item response data in education. However, it is also very restrictive in that it assumes that examinees need to possess all the attributes required for an item to answer it correctly, and the absence of a single required attribute is not different from the absence of all the required attributes. The G-DINA model was proposed to relax this specific assumption and allows examinees with various subsets of the required attributes to have different probabilities of success on the item. Instead of two latent groups, the G-DINA model partitions the latent classes into $2^{K_j^*}$ latent groups, where $K_j^* = \sum_{k=1}^{K} q_{jk}$ represents the number of required attributes for item j and q_{jk} is the kth element of the jth row of the Q-matrix (K. K. Tatsuoka, 1983). For notational convenience but without loss of generality, let attributes $1, \ldots, K_j^*$ be required for item j and $\boldsymbol{\alpha}_{lj}^*$ be the reduced attribute vector consisting of the columns of the required attributes, where $l = 1, \ldots, 2^{K_j^*}$. The probability that examinees with attribute pattern $\boldsymbol{\alpha}_{lj}^*$ will answer item j correctly is denoted by $P(X_j = 1|\boldsymbol{\alpha}_{lj}^*) = P(\boldsymbol{\alpha}_{lj}^*)$. The probability of success on item j given the reduced attribute vector $\boldsymbol{\alpha}_{lj}^*$ also can be written as

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk} + \sum_{k'>k}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{jkk'}\alpha_{lk}\alpha_{lk'} \ldots + \delta_{j12\cdots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{lk}. \quad (1)$$

As can be seen from (1), the original formulation of the G-DINA model based on $P(\boldsymbol{\alpha}_{lj}^*)$ can be reparameterized in terms of $\delta_{j\cdot}$, where δ_{j0} is the intercept for item j, δ_{jk} is the main effect due to α_{jk}, $\delta_{jkk'}$ is the interaction effect due to α_{jk} and $\alpha_{jk'}$, and $\delta_{j12\ldots K_j^*}$ is the interaction effect due to $\alpha_1, \ldots, \alpha_{K_j^*}$.

Several specific models can be derived from the G-DINA model. The DINA model can be obtained from the G-DINA model by setting all the parameters—except δ_0 and $\delta_{12\ldots K_j^*}$—to zero, whereas the DINO model can be obtained by setting

$$\delta_{jk} = -\delta_{jk'k''} = \cdots = (-1)^{K_j^*+1}\delta_{j12\ldots K_j^*},$$

for $k = 1, \ldots, K_j^*$, $k' = 1, \ldots, K_j^* - 1$, and $k'' > k', \ldots, K_j^*$. The additive CDM (*A*-CDM) can be obtained from the G-DINA model by setting all the interaction effects to zero. That is,

$$P(\boldsymbol{\alpha}_{lj}^*) = \delta_{j0} + \sum_{k=1}^{K_j^*} \delta_{jk}\alpha_{lk}. \tag{2}$$

The G-DINA model is based on the identity link function, and is equivalent to other general CDMs based on alternative link functions (i.e., logit, log). Incidentally, by applying the same constraints, the DINA and DINO models can also be derived from the logit and log CDMs, albeit on different scales. In addition, by setting the interaction terms to zero, the linear logistic model (LLM; Hagenaars, 1990, 1993; Maris, 1999) and the R-RUM can be obtained from the logit and log CDMs, respectively. When the *noisy inputs, deterministic "and" gate* (Junker & Sijtsma, 2001) model is generalized to allow the slip and guessing parameters to vary across items, the resulting model can also be shown to be the additive log CDM.

Estimation

As shown by de la Torre (2011), the marginal maximum likelihood estimate (MMLE) of the parameter $P(\boldsymbol{\alpha}_{lj}^*)$ using an EM algorithm implementation is given by

$$\hat{P}(\boldsymbol{\alpha}_{lj}^*) = \frac{R_{\boldsymbol{\alpha}_{lj}^*}}{I_{\boldsymbol{\alpha}_{lj}^*}}, \tag{3}$$

where $I_{\boldsymbol{\alpha}_{lj}^*}$ is the number of examinees expected to be in the latent group $\boldsymbol{\alpha}_{lj}^*$, and $R_{\boldsymbol{\alpha}_{lj}^*}$ is the number of examinees in group $\boldsymbol{\alpha}_{lj}^*$ expected to answer item j correctly. An approximation of the standard error of the estimate, $SE[\hat{P}(\boldsymbol{\alpha}_{lj}^*)]$, can be obtained from the information matrix. De la Torre noted that the algorithm for estimating the G-DINA model parameters and the corresponding SEs largely is similar to the EM algorithms in estimating the parameters of the DINA and multiple-choice DINA models described in detail by de la Torre (2009a, 2009b). The parameter estimates of the reduced CDMs described above can be derived efficiently from the $2^{K_j^*}$ G-DINA model parameter estimates (de la Torre & Chen, 2010).

The Wald Test for the G-DINA Model

In presenting the G-DINA model, de la Torre (2011) also proposed the Wald test as an item-level statistical test to examine whether one of the interpretable reduced models can be used in place of the saturated G-DINA model without adversely affecting the model-data fit. The Wald test is an item-level procedure in that it can be performed one item at a time whenever $K_j^* > 1$. The following sections describe how the Wald test can be carried out in the context of the G-DINA model.

To examine the adequacy of a reduced model, the Wald test requires setting up a $(2^{K_j^*} - p) \times 2^{K_j^*}$ restriction matrix $\boldsymbol{R}$, where p represents the number of parameters of the reduced model. Under specific constraints, the saturated model is equivalent to the reduced model of interest. For example, when $K_j^* = 3$, the DINA model, the DINO model, and the *A*-CDM have $p = 2$, 2, and 4 parameters, respectively. The corresponding restriction matrices for these reduced models are:

$$\boldsymbol{R}_{6\times 8}^{(1)} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \end{pmatrix}, \tag{4}$$

$$\boldsymbol{R}_{6\times 8}^{(2)} = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}, \tag{5}$$

and

$$\boldsymbol{R}_{4\times 8}^{(3)} = \begin{pmatrix} 1 & -1 & -1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & -1 & 0 & -1 & 0 & 0 \\ 1 & 0 & -1 & -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \end{pmatrix}. \tag{6}$$

The restriction matrices $\boldsymbol{R}^{(1)}$, $\boldsymbol{R}^{(2)}$, and $\boldsymbol{R}^{(3)}$ yield the following sets of constraints:

$$\delta_1 = \delta_2 = \delta_3 = \delta_{12} = \delta_{13} = \delta_{23} = 0,$$

$$\delta_1 = \delta_2 = \delta_3 = -\delta_{12} = -\delta_{13} = -\delta_{23} = \delta_{123}, \quad \text{and}$$

$$\delta_{12} = \delta_{13} = \delta_{23} = \delta_{123} = 0,$$

respectively. These equalities show that when $K_j^* = 3$, the main effects and two-way interactions—but not the intercept and three-way interaction—are all equal to zero for the DINA model; main effects and three-way interaction are equal to each other and the negative of the two-way interactions for the DINO model; and all the interaction terms are equal to zero for the *A*-CDM.

Table 1
Probability of Success for a Two-Attribute Item for the Five Generating Models

	Latent Group			
Generating Model	00	10	01	11
DINA	.20	.20	.20	.80
DINA/*A*-CDM	.20	.35	.35	.80
A-CDM	.20	.50	.50	.80
DINO/*A*-CDM	.20	.65	.65	.80
DINO	.20	.80	.80	.80

The Wald statistic W is then computed as

$$W = [\boldsymbol{R} \times \boldsymbol{P}_j]'\{\boldsymbol{R} \times \text{Var}(\boldsymbol{P}_j) \times \boldsymbol{R}'\}^{-1}[\boldsymbol{R} \times \boldsymbol{P}_j], \tag{7}$$

where $\boldsymbol{P}_j = \{P(\boldsymbol{\alpha}_{lj}^*)\}$ and $\text{Var}(\boldsymbol{P}_j)$ is the inverse of the information matrix. Under the null hypothesis that $\boldsymbol{R} \times \boldsymbol{P}_j = \boldsymbol{0}$, W is assumed to be asymptotically χ^2-distributed with $2^{K_j} - p$ degrees of freedom. To implement the Wald test, $\boldsymbol{P}_j$ and $\text{Var}(\boldsymbol{P}_j)$ are replaced by their sample counterparts, as in, $\hat{\boldsymbol{P}}_j$ and $\text{Var}(\hat{\boldsymbol{P}}_j)$.

Compared to model selection at the test level, the Wald test for the G-DINA model does not require blanket acceptance of a model or possibly a collection of models for all items comprising the test. Such a process can lead to suboptimal choices. Instead, the test affords researchers the flexibility of investigating the CDM that best fits each of the multiattribute items.

Current Study

Notwithstanding its potential usefulness, the Wald test cannot be adopted for practical use without a clearer understanding of its statistical properties. Thus, the primary goal of this work was to use a simulation study to systematically examine the Type I error and power of the Wald test in the context of the G-DINA model. Specifically, these properties will be documented for three reduced forms of the G-DINA model—the DINA model, the DINO model, and the *A*-CDM. In addition to a simulation study, we also examined the performance of the Wald test with real data.

Simulation Study

Design. In documenting the Type I error and power of the Wald test, three factors were examined: sample size, focal model, and generating model. The Type I component of the study focused on three reduced models (i.e., DINA, DINO, and *A*-CDM). For the power component of the study, each of the focal models was used to analyze data generated from other models. In addition to the focal models, data also were generated from a model derived from the combination of the DINA and *A*-CDM and a model derived from the combination of the DINO and *A*-CDM. Table 1 illustrates the similarities and differences of the five generating models, and gives the probability of success of each latent group when an item requires two attributes under the identity link. The same models are represented graphically in Figure 1.

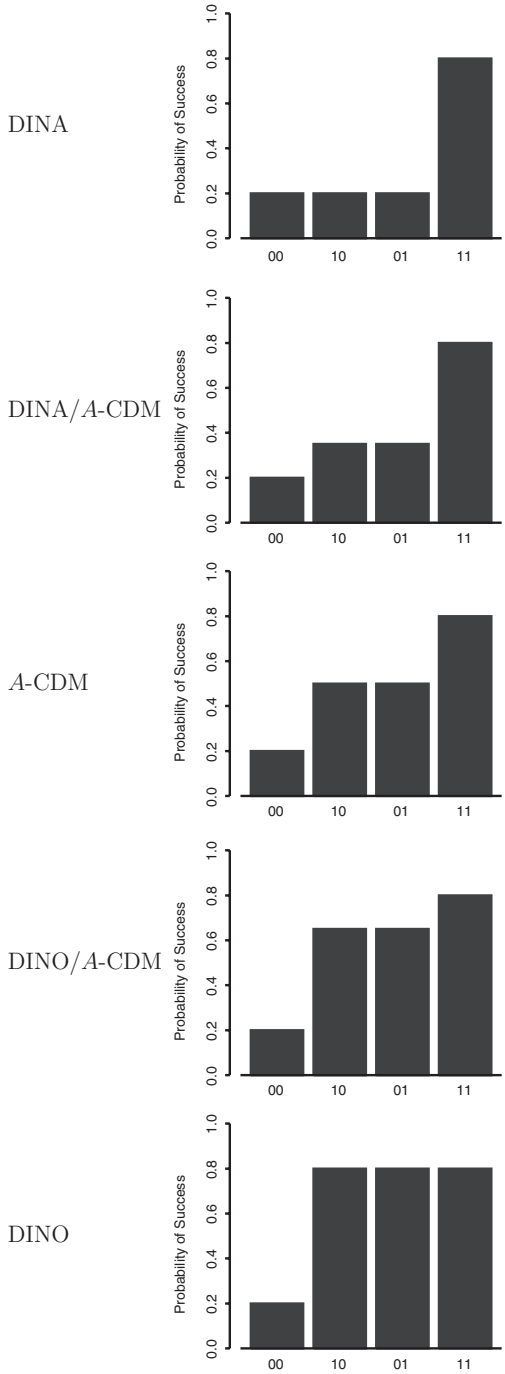

Figure 1. Success probabilities for various instances of the G-DINA model when $K_j^* = 2$.

Table 2
Simulation Study Q-Matrix

Item	Attribute					Item	Attribute				
	α_1	α_2	α_3	α_4	α_5		α_1	α_2	α_3	α_4	α_5
1	1	0	0	0	0	16	0	1	0	1	0
2	0	1	0	0	0	17	0	1	0	0	1
3	0	0	1	0	0	18	0	0	1	1	0
4	0	0	0	1	0	19	0	0	1	0	1
5	0	0	0	0	1	20	0	0	0	1	1
6	1	0	0	0	0	21	1	1	1	0	0
7	0	1	0	0	0	22	1	1	0	1	0
8	0	0	1	0	0	23	1	1	0	0	1
9	0	0	0	1	0	24	1	0	1	1	0
10	0	0	0	0	1	25	1	0	1	0	1
11	1	1	0	0	0	26	1	0	0	1	1
12	1	0	1	0	0	27	0	1	1	1	0
13	1	0	0	1	0	28	0	1	1	0	1
14	1	0	0	0	1	29	0	1	0	1	1
15	0	1	1	0	0	30	0	0	1	1	1

The layout and ordering of the graphs indicate which models are considered more similar to or dissimilar from each other. For example, the DINA and DINO models are the two most different from each other, and the *A*-CDM is more similar to DINA/*A*-CDM or DINO/*A*-CDM than it is to either the DINA or DINO model. Irrespective of the model and number of required attributes for an item, the probabilities of success for individuals who mastered none and all of the required attributes were fixed to .20 and .80, respectively. For the additive model, an increment of $.60/K_j^*$ is associated with each attribute mastery. Finally, the probabilities of success for the DINA/*A*-CDM are obtained by averaging the DINA and *A*-CDM probabilities of success associated with the same latent group.

The three sample sizes in the simulation study were $I = 500$, 1,000, and 2,000. The numbers of items and attributes were fixed to $J = 30$ and $K = 5$, respectively, and the attribute vectors were generated from a uniform distribution. The Q-matrix used in simulating the response data is given in Table 2. For each sample size-model combination, 1,000 data sets were generated and analyzed. The Type I error and power of the Wald test were investigated using nine significance levels: .50, .40, .30, .20, .10, .05, .02, .01, and .005. The estimation code used in this article was written in Ox (Doornik, 2003). The code can be made available by contacting the authors.

Results.

Type I error. Table 3 gives the Type I error of the DINA model, the DINO model, and the *A*-CDM. Due to the symmetry of the design, results were averaged within an item type (i.e., item with the same number of required attributes) and across δ_j of the same value when applicable (i.e., $K_j^* = 2$ or 3 for the *A*-CDM). The Type I errors for

Table 3
Type I Error of the Wald Test for the Three Reduced Models

		Observed Proportions					
		DINA		*A*-CDM		DINO	
I	α	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$
	.500	.524	.560	.517	.544	.517	.560
	.400	.423	.471	.419	.450	.418	.476
	.300	.327	.377	.208	.351	.323	.383
	.200	.222	.280	.216	.247	.221	.275
500	.100	.116	.163	.114	.137	.120	.167
	.050	.062	.099	.059	.076	.066	.104
	.020	.029	.050	.023	.035	.030	.055
	.010	.015	.030	.012	.020	.018	.033
	.005	.009	.018	.007	.012	.011	.019
	.500	.505	.537	.499	.519	.508	.540
	.400	.407	.441	.402	.421	.409	.443
	.300	.311	.345	.194	.318	.311	.341
	.200	.211	.248	.201	.218	.210	.240
1,000	.100	.108	.139	.102	.112	.111	.137
	.050	.059	.085	.054	.059	.061	.082
	.020	.024	.044	.021	.025	.026	.047
	.010	.013	.028	.011	.015	.014	.030
	.005	.008	.020	.005	.008	.008	.020
	.500	.510	.521	.507	.512	.507	.520
	.400	.412	.418	.408	.413	.404	.427
	.300	.304	.324	.199	.313	.302	.327
	.200	.201	.223	.208	.212	.200	.224
2,000	.100	.100	.119	.107	.112	.104	.119
	.050	.049	.064	.052	.062	.054	.066
	.020	.020	.029	.020	.025	.021	.031
	.010	.010	.017	.010	.013	.011	.018
	.005	.009	.018	.006	.006	.006	.011

$K_j^* = 2$ and $K_j^* = 3$ are documented separately. Figure 2 graphically represents the Type I error of the three reduced models. In these figures, $W2$ and $W3$ denote $K_j^* = 2$ and $K_j^* = 3$, respectively. The leftmost panels of the figure show that the Type I error of the Wald test for the DINA model overestimated the nominal significance level across all conditions. However, the overestimation became less apparent when $K_j^* = 2$ and the sample size was larger. The same pattern can be observed for the DINO model (see rightmost panels). A different pattern can be observed for the Type I error rate of the *A*-CDM. When the Wald test was used in conjunction with the *A*-CDM, the Type I error rates were closer to the nominal significance levels compared to those for the DINA or DINO models (see middle panels). Overall, the Type I error rates for the three models were comparable and close to the nominal value when the

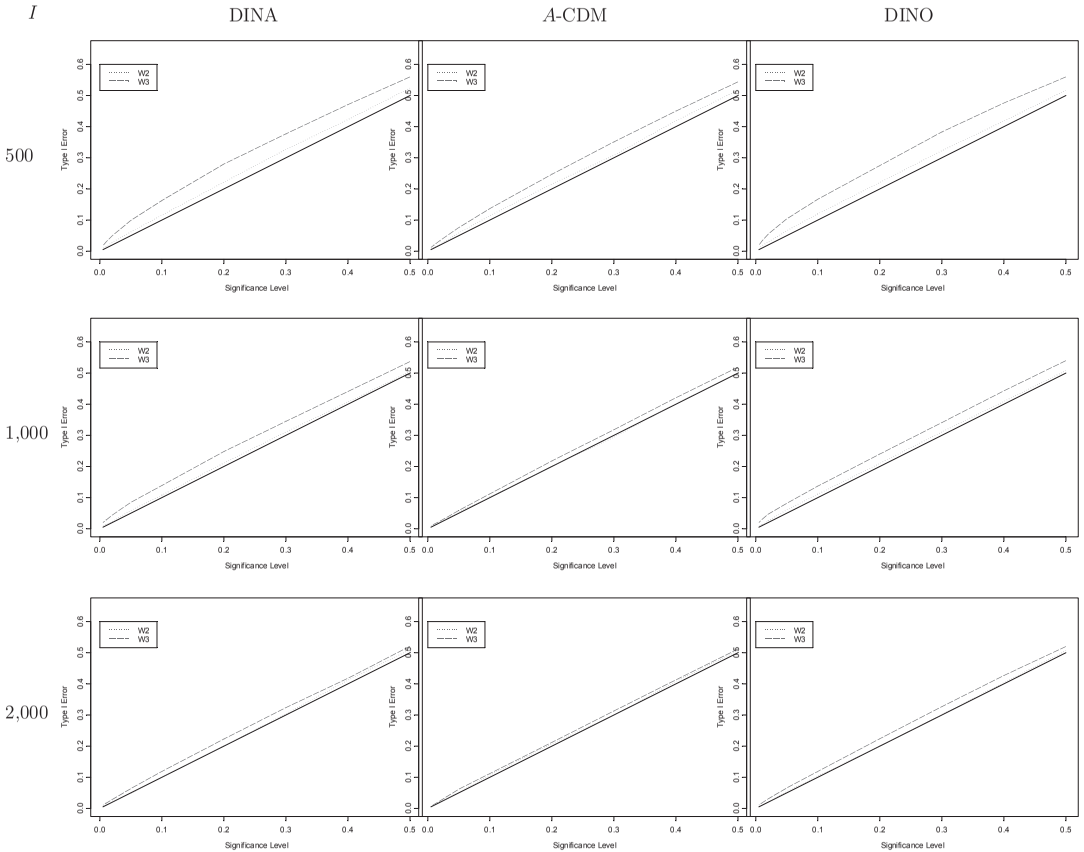

Figure 2. Type I error of the Wald test for the three reduced models.

sample size was large. For example, at a significance level of .05 and $I = 2,000$, the Type I error rates of the Wald test for the DINA model, the DINO model, and the A-CDM were between .049 and .066 for both $K_j^* = 2$ and 3.

Power. In the power analysis literature, a test power of at least .80 is considered adequate. For the purposes of this article, we also define a test power greater than or equal to .90 as excellent. The previous section indicates that comparing the test statistics to the theoretical χ^2 distributions can lead to inflated Type I error. Consequently, without any correction, the power of the Wald test also will be overestimated. For this reason, in addition to the theoretical distributions we also use the quantiles from the empirical distributions of the test statistic to determine acceptance or rejection of the null hypotheses. Again, because the saturated model cannot be reduced when $K_j^* = 1$, tests were performed only for items where $K_j^* = 2$ and $K_j^* = 3$.

Tables 4 through 7 display the power of the Wald test based on both the empirical and theoretical distributions. It should be noted that each of the reduced models (i.e., DINA, DINO, and A-CDM) was examined using data from four other generating models. Of the 12 possible tables, only four tables where power is less than 1.00 are presented. For example, only when data were generated, using the DINA/A-CDM was the power of the Wald tests less than perfect for the DINA model (see Table 4). This is because the DINA/A-CDM was the generating model most similar to the DINA model. The remaining generating models (i.e., A-CDM, DINO/A-CDM, and DINO) were sufficiently different from the DINA model; hence, the tests consistently flagged each multiattribute item as not conforming to the DINA model (i.e., the power was equal to 1.0).

A closer inspection of Table 4 indicates that the power of the Wald test based on the theoretical distribution was excellent except for conditions with $K_j^* = 3$ at the nominal significance level of .010 and .005, and $I = 500$. Under these conditions, the power of the test ranged from .81 to .86 and thus remained adequate. However, as expected, when corrections were made by using the empirical distributions, inadequate test powers were obtained for these conditions. In addition, for $K_j^* = 2$ at a significance level of .005, or for $K_j^* = 3$ and a significance level less than or equal to .02, less than excellent test powers were obtained. The use of theoretical and empirical cutoffs showed the conditions under which the theoretical and empirical power rates can differ substantially. For example, the largest differences between the power rates were observed for the DINA and DINO models when data were generated using the DINA/A-CDM and DINO/A-CDM, respectively, $K_j^* = 3$, $I = 500$, and the significance level was less than .05; for the remaining conditions, the theoretical and empirical power rates were comparable. Finally, as expected, the simulation results showed that better powers were obtained with larger sample sizes.

Table 5 shows the power of the Wald test for the A-CDM when the data were generated using the DINA/A-CDM. All conditions resulted in test powers that were excellent to perfect. The same pattern was observed when fitting the A-CDM to the data generated using the DINO/A-CDM (see Table 6). Lastly, in fitting the DINO model to the data generated using the DINO/A-CDM, the results in Table 7 show powers very similar in magnitude to the powers obtained in fitting the DINA model to the data generated using the DINA/A-CDM.

Table 4
Power of the Wald Test for the DINA Model: DINA/A-CDM-Generated Data

		Theoretical		Empirical	
I	α	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	.99
	.300	1.00	.99	1.00	.99
	.200	1.00	.99	1.00	.98
500	.100	.99	.97	.99	.95
	.050	.98	.95	.97	.90
	.020	.96	.90	.94	.82
	.010	.93	.86	.91	.74
	.005	.90	.81	.85	.66
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
1,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	.99
	.005	1.00	1.00	1.00	.98
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
2,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00

Real Data Example

Fraction subtraction data. To provide valuable information concerning the practical utility of the Wald test, the behavior of the test was examined using fraction subtraction data described and used by K. K. Tatsuoka (1990) and more recently originally by C. Tatsuoka (2002, 2005) and de la Torre (2011). As in de la Torre (2011), only a subset of these data were used in this example. These data represented the responses of 536 middle school students to 12 fraction subtraction problems involving four attributes: (a) performing basic fraction subtraction operations, (b) simplifying/reducing, (c) separating whole numbers from fractions, and (d) borrowing one from whole numbers to fractions. The fraction subtraction items and the corresponding required attributes are given in Table 8.

Table 5
Power of the Wald Test for the A-CDM: DINA/A-CDM-Generated Data

		Theoretical		Empirical	
I	α	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
500	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	.99
	.020	1.00	.99	1.00	.98
	.010	.99	.98	.99	.96
	.005	.99	.96	.99	.94
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
1,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
2,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00

The Wald test was carried out for the 10 multiattribute items in the test (i.e., all but Items 1 and 3). For these items, three reduced models—DINA, DINO, and *A*-CDM—were considered. As with the original G-DINA model analysis of these data, the constraint $P(\boldsymbol{\alpha}_{lj}^*) \leq P(\boldsymbol{\alpha}_{l'j}^*)$ if $\alpha_{ljk} \leq \alpha_{l'jk}$ for $l \neq l'$ and $k = 1, \ldots, K_j^*$ also was imposed to stabilize the estimates.

Results. The Wald statistics W and p-values for all the tests are given in Table 9. The results showed that at $\alpha = .05$, either the DINA model or the *A*-CDM can adequately fit the fraction subtraction items; however, the DINO model cannot adequately fit any of the items. The DINA model can be used with items 2, 5, 6, and 10 and the *A*-CDM with items 7, 8, 9, 11, and 12, whereas both models can be used with item 4. However, it can be argued based on the p-values that the *A*-CDM is

Table 6
Power of the Wald Test for the A-CDM: DINO/A-CDM-Generated Data

		Theoretical		Empirical	
I	α	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
500	.100	1.00	1.00	1.00	1.00
	.050	1.00	.99	1.00	.99
	.020	1.00	.98	1.00	.97
	.010	1.00	.97	.99	.96
	.005	.99	.96	.99	.94
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
1,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
2,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00

more appropriate for item 4. When the significance level is changed to $\alpha = .1$, the most appropriate reduced model for each item becomes apparent.

Discussion

With the availability of multiple viable CDMs, objectively choosing the most appropriate model is crucial. This article examined the Wald test for comparing saturated and reduced models under the G-DINA framework. The simulation study shows that the Wald test is promising in that it provides relatively accurate Type I error when the sample size is large and the number of parameters is small. The simulation study also shows that when reasonable significance levels are involved (e.g., .05), the Wald test has excellent power to detect that the true underlying model is neither DINA, DINO, or *A*-CDM even for relatively small sample sizes.

Table 7
Power of the Wald Test for the DINO Model: DINO/A-CDM-Generated Data

		Theoretical		Empirical	
I	α	$K_j^* = 2$	$K_j^* = 3$	$K_j^* = 2$	$K_j^* = 3$
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	.99
	.300	1.00	.99	1.00	.99
	.200	1.00	.99	1.00	.98
500	.100	.99	.97	.99	.94
	.050	.98	.94	.97	.89
	.020	.95	.90	.93	.82
	.010	.93	.86	.89	.76
	.005	.90	.82	.84	.71
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
1,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	.99
	.010	1.00	1.00	1.00	.98
	.005	1.00	.99	1.00	.97
	.500	1.00	1.00	1.00	1.00
	.400	1.00	1.00	1.00	1.00
	.300	1.00	1.00	1.00	1.00
	.200	1.00	1.00	1.00	1.00
2,000	.100	1.00	1.00	1.00	1.00
	.050	1.00	1.00	1.00	1.00
	.020	1.00	1.00	1.00	1.00
	.010	1.00	1.00	1.00	1.00
	.005	1.00	1.00	1.00	1.00

Table 8
Fraction Subtraction Items and Required Attributes

		Attribute						Attribute			
Item		1	2	3	4	Item		1	2	3	4
1.	$\frac{3}{4} - \frac{3}{8}$	1	0	0	0	7.	$\frac{11}{8} - \frac{1}{8}$	1	1	0	0
2.	$3\frac{1}{2} - 2\frac{3}{2}$	1	1	1	1	8.	$3\frac{4}{5} - 3\frac{2}{5}$	1	0	1	0
3.	$\frac{6}{7} - \frac{4}{7}$	1	0	0	0	9.	$4\frac{5}{7} - 1\frac{4}{7}$	1	0	1	0
4.	$3\frac{7}{8} - 2$	1	0	1	0	10.	$7\frac{3}{5} - \frac{4}{5}$	1	0	1	1
5.	$4\frac{4}{12} - 2\frac{7}{12}$	1	1	1	1	11.	$4\frac{1}{10} - 2\frac{8}{10}$	1	1	1	1
6.	$4\frac{1}{3} - 2\frac{4}{3}$	1	1	1	1	12.	$4\frac{1}{3} - 1\frac{5}{3}$	1	1	1	1

Table 9
W and p-Value of the Wald Test for Multiattribute Fraction Subtraction Items

	DINA		DINO		*A*-CDM	
Item	*W*	*p*	*W*	*p*	*W*	*p*
2	5.78	.972	346.00	.000	27.25	.004
4	5.92	.052	13.84	.001	.07	.787
5	19.00	.165	580.71	.000	25.60	.007
6	16.42	.288	1,442.30	.000	45.80	.000
7	38.37	.000	45.08	.000	.09	.761
8	46.43	.000	48.06	.000	.00	.998
9	52.62	.000	42.26	.000	1.58	.208
10	4.98	.547	866.74	.000	23.30	.000
11	48.23	.000	344.45	.000	3.88	.973
12	29.29	.009	630.37	.000	13.43	.267

This article provides an important contribution toward advancing the use of cognitive diagnosis modeling in practical educational settings. The G-DINA model subsumes several commonly used CDMs and, with the Wald test statistically determining item by item whether a reduced model can be used in place of the saturated model, becomes feasible. This can make the application of cognitive diagnosis modeling more flexible in that (a) CDMs need not be specified *a priori*, and (b) multiple, statistically determined CDMs can be used within a single assessment.

Despite these advances, it is important to note that the results from the current study further underscore the need to investigate how model comparison at the item level can be done more reliably. It would be useful to explore whether alternative statistics that produce more accurate Type I error rates without substantial loss of test power can be found. This will obviate the need to base decisions on empirical distributions that can be time-consuming to document when considering different scenarios such as type of reduced CDM, sample size, number of attributes, test length, and item quality. If the proposed test is to be retained, it would be helpful to determine whether corrections that adjust for model complexity such as the AIC (Aikake, 1973) can be incorporated into the current formulations of the Wald test.

It should be noted that the Wald test as used in this article only examined whether the saturated model can be simplified into one or more of the reduced models. However, the current application of the test did not directly examine whether other parts of the model (e.g., attribute specifications) can be simplified. For greater generality, future simulation studies should include more varied conditions. Among other factors, it would be instructive to examine how attribute distribution, discrepancy between the focal and generating models, and specification and completeness of the Q-matrices affect the Type I error and power of the test. Exploring how the test can be adapted to reduced models of other link functions (e.g., additive models under the log and logit links) also will engender confidence in the generality of the proposed method.

It is also worth noting that although the primary goal of cognitive diagnosis modeling is to provide scores (i.e., estimates of attribute masteries) that can be used to inform instruction and learning, CDMs can provide other information that can have theoretical and practical implications. In this study, we focused on determining how the attributes relate to individual items. As such, the specific form of the CDM for an item indicates the type of interaction present between the required attributes and the item. For example, if the DINA model is deemed appropriate for the item, it can be inferred that students need to demonstrate mastery of all the required attributes to answer the item correctly. On the other hand, if the DINO model is deemed appropriate, then students can be given the option to choose which of the required attributes to master in order to answer the item correctly. Similarly, the specific form of the CDM can be used to examine the features of the items that make them conjunctive, disjunctive, additive, or otherwise. This type of analysis would be particularly interesting if items that require the same set of attributes but have different attribute-item interactions exist.

Finally, as with traditional models, caution needs to be exercised in interpreting and using scores derived from CDMs. Effort needs to be expended in ensuring that relevant attributes are defined, appropriate assessment items are used, pertinent pieces of evidence are gathered, and representative members of the population of interest are sampled before employing test scores from CDMs as the basis for making decisions, particularly if these decisions can have serious consequences.

Acknowledgments

This research was supported by National Science Foundation Grant DRL-0744486.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proceedings of the Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akad. Kiado.

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.

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix. *Applied Psychological Measurement*, *35*, 8–26.

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively-based multiple-choice options. *Applied Psychological Measurement*, *33*, 163–183.

de la Torre, J. (2009b). 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., & Chen, J. (2010, April). *Estimating different reduced cognitive diagnosis models using a general framework*. Paper presented at the meeting of the National Council on Measurement in Education, New Orleans, LA.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

de la Torre, J., & Karelitz, T. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure. *Journal of Educational Measurement*, *46*, 450–469.

DiBello, L., Roussos, L., & Stout, W. (2007). Cognitive diagnosis Part I. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics*. Amsterdam, The Netherlands: Elsevier.

Doornik, J. A. (2003). *Object-oriented matrix programming using Ox (Version 3.1)*. [Computer software]. London, UK: Timberlake Consultants Press.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 333–352.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Loglinear panel, trend, and cohort analysis*. Thousand Oaks, CA: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Thousand Oaks, CA: Sage.

Hambleton, R. K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In W. R. Lenderking & R. Reviki (Eds.), *Advances in health outcomes research methods, measurement, statistical analysis, and clinical applications* (pp. 57–78). Washington, DC: Degnon Associates.

Hartz, S. (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.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *44*, 191–210.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

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.

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

Rojas, G., de la Torre, J., & Olea, J. (2012, April). *Choosing between general and specific cognitive diagnosis models when the sample size is small*. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

Schwarz, G. (1976). Estimating the dimension of a model. *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.

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*, *64*, 583–639.

Swaminathan, H., Hambleton, R. K., & Rogers, H. J. (2007). Assessing the fit of item response theory models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics*, Vol. *26* (pp. 683-718). New York, NY: Elsevier.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Applied Statistics*, *51*, 337–350.

Tatsuoka, C. (2005). Corrigendum: Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*, 465–467.

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

Tatsuoka, K. K. (1990). *Toward an integration of item-response theory and cognitive error diagnosis*. In N. Fredrickson, R. L. Glaser, A. M. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum.

Templin, J. (2006). *DCM user's guide*. Lawrence: University of Kansas.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report 05-16). Princeton, NJ: Educational Testing Service.

Authors

JIMMY DE LA TORRE is Associate Professor of Educational Psychology at Rutgers University, 10 Seminary Place, New Brunswick, NJ, 08901; e-mail: j.delatorre@rutgers.edu. His primary research interests include item response theory, cognitive diagnosis, Bayesian analysis, and the use of diagnostic assessments to support classroom instruction and learning.

YOUNG-SUN LEE is Associate Professor of Psychology and Education, Teachers College–Columbia University, 525 West 120th Street, New York, NY 10027; e-mail: yslee@tc.columbia.edu. Her primary research interests include applications of item response theory and cognitive diagnosis modeling.