Article

Journal of Educational and Behavioral Statistics
2017, Vol. 42, No. 6, pp. 651–677
DOI: 10.3102/1076998617703060

A Cognitive Diagnosis Model for Continuous Response

Nathan D. Minchen
Rutgers, The State University of New Jersey

Jimmy de la Torre
The University of Hong Kong

Ying Liu
University of Southern California

Nondichotomous response models have been of greater interest in recent years due to the increasing use of different scoring methods and various performance measures. As an important alternative to dichotomous scoring, the use of continuous response formats has been found in the literature. To assess finer-grained skills or attributes and to extract information with diagnostic value from continuous response data, a multidimensional skills diagnosis model for continuous response is proposed. An expectation-maximization implementation of marginal maximum likelihood estimation is developed to estimate its parameters. The viability of the proposed model is shown via a simulation study and a real data example. The proposed model is also shown to provide a substantial improvement in attribute classification when compared to a model based on dichotomized continuous responses.

Keywords: *cognitive diagnosis models; continuous response; DINA model*

1. Introduction

Due to the increasing use of different scoring methods and various performance measures, interest in nondichotomous response models has grown in the last several decades. Recent research, some of which will be discussed in this article, has addressed polytomous item response models in the context of both traditional item response theory (IRT) and cognitive diagnosis models (CDMs). However, far less effort, particularly for CDMs, has been devoted to modeling continuous response, which has been found in at least the following three areas and will be discussed in turn: (1) providing a level of endorsement by marking a continuum, (2) "probability testing," and (3) the use of latency data.

The purpose of this article is to present an adaptation of the deterministic inputs, noisy "AND" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001) model that handles continuous response. Although this model is presented as a general framework that may be used or modified for other types of continuous response measures, we will discuss it in the context of analyzing latency data. In many cases, latency data are readily available alongside response accuracy data, and proper analysis may yield additional diagnostic insights. The sections that follow will discuss the aforementioned three types of continuous response data, with a brief overview of some of the models that have been developed to analyze such data. Next, a review of CDMs will be presented, from which the continuous response version of the DINA model and its estimation procedures will be developed. Finally, a simulation study and a real data example, followed by a discussion, will demonstrate the viability of the proposed model.

1.1. Continuous Response as a Measure

In some tasks that appear on personality and attitude assessments, respondents are asked to report their level of endorsement with various statements by placing a mark somewhere on a line segment, the ends of which represent the lack of endorsement and complete endorsement. The distance from one of the ends of the segment to the mark can be used as the measure. Noel and Dauvier (2007) have developed an IRT model to handle such a response. Their model is predicated on the notion that such a response format is essentially an infinite-category Likert-type scale. They discuss the idea that more Likert categories are preferable to fewer categories because they provide finer-grained measurement but that the downside of additional categories is that they require the estimation of additional parameters. In general, a Likert-type scale of a large number of graded options is usually considered to be continuous (e.g., Samejima, 1973; Thissen, Steinberg, Pyszczynski, & Greenberg, 1983). Noel and Dauvier (2007) postulate that the β distribution has properties that make it an appropriate choice for the "interpolation response mechanism" that this type of response measure represents (p. 49). An extension of this model by Noel (2014) can handle responses that are of an unfolding nature. This scoring method can also be used in other situations in which a Likert-type scale would be appropriate, such as pain intensity assessment (e.g., Morin & Bushnell, 1998).

Another source of continuous response that is similar to the response format in the previous paragraph comes from probability testing. Using multiple-choice questions, examinees are asked to report the *probability* that each option is the correct answer rather than actually choosing one of the alternatives. These probabilities are considered to reveal partial knowledge (de Finetti, 1965). In some situations, this method is simplified. In one such simplification, for example, the examinees are asked to express their confidence only for the most correct option, usually by using a Likert-type scale. Such a scoring method is referred to as

"confidence marking" and was, to our knowledge, first suggested and applied by Dressel and Schmidt (1953, as cited in Ben-Simon, Budescu, & Nevo, 1997). The merits of both methods over dichotomous scoring are discussed by Ben-Simon, Budescu, and Nevo (1997).

Perhaps one of the most popular sources of continuous responses is response time, thanks in large part to the rapid expansion of computer-based testing in recent years. In computer-based testing, response times have been used to detect aberrant responses (van der Linden & Guo, 2008; van der Linden & van Krimpen-Stoop, 2003), improve item selection in computerized adaptive testing (Fan, Wang, Chang, & Douglas, 2012; Sie, Finkelman, Riley, & Smits, 2015; van der Linden, 2008), control differential speededness in computerized adaptive testing (van der Linden, 2009; van der Linden, Scrams, & Schnipke, 1999; van der Linden & Xiong, 2013), and control differential speededness in multistage testing (van der Linden, Breithaupt, Chuah, & Zhang, 2007). Response times have also been used to improve parameter estimates (Meng, Tao, & Chang, 2015; Ranger & Kuhn, 2012), ability estimates (Ferrando & Lorenzo-Seva, 2007 Meng et al., 2015; Meng, Tao, & Shi, 2014), and ability classifications (Sie et al., 2015) under certain conditions. Uncertainty on personality assessments can also be assessed using response time (e.g., Ferrando & Lorenzo-Seva, 2007; Meng et al., 2014). Finally, the fact that most power tests are administered under time constraints necessitates the study of latency in addition to correctness (e.g., Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997).

To address the growing prevalence and importance of continuous measures, a number of continuous response models have been proposed for the IRT framework. Possibly the earliest development of a continuous response model in IRT is by Samejima (1973), who showed that the "limiting" (p. 204) case of the graded response model (Samejima, 1969) can be viewed as a continuous response model. Such a response model can be used either in the unidimensional (Samejima, 1973), for which Wang and Zeng (1998) have developed an expectation-maximization (EM) estimation algorithm, or in the multidimensional case (Samejima, 1974).

More recently, van der Linden (2007) has provided a model framework for studying response time and response accuracy simultaneously. In the first level of his hierarchical model, separate item response models for response time and response accuracy are specified, each of which arises from a unique latent trait. The second level of the model relates latent traits through a multivariate normal distribution. Van der Maas and Wagenmakers (2005) use response time in a different way. They develop a model in which the scoring rule itself incorporates response time, rather than dealing with response time separately. Total test score is comprised of the sum of the time remaining for each question (i.e., the time limit less the response time) summed over all correct items. Maris and van der Maas (2012) note that such a scoring rule may encourage guessing behavior and propose an adjusted scoring rule that penalizes incorrect responses that are made too quickly.

It is important to note that these models all assume that the underlying latent trait is continuous, as is typical in IRT, with the exception of van der Linden's (2007) framework, which is less specific in the sense that it can handle any item response model. As such, there is a need for continuous response models in the cognitive diagnosis modeling framework.

1.2. CDMs

In a typical IRT application, the continuous latent trait is usually unidimensional and broadly defined. In contrast, CDMs use a latent, multidimensional, discrete vector as the person-specific variable and are designed to assess finer-grained skills and to extract information with more diagnostic value. The goal of analyzing data with a CDM is generally to determine which of a set of discrete skills examinees have mastered. The set of skills that examinee i possesses is referred to as the attribute pattern, which is an unobservable latent variable represented by a vector of length K, denoted as α_i. Entries of 1 and 0 indicate the presence or absence of that skill, respectively. Correspondingly, a **q**-vector, denoted as $\mathbf{q}_j$, specifies which of the one or more of the K skills are measured by item j. CDMs are further defined by the way in which they conceptualize the interaction between an examinee's skills vector and the item's **q**-vector. For some CDMs, this interaction can be formalized in the definition of the *latent response variable.* Several well-known CDMs are discussed next.

The DINA (Haertel, 1989; Junker & Sijtsma, 2001) model is a commonly used and readily interpretable CDM. Its latent response variable is defined such that only examinees who possess all of the skills required to solve a given item are expected to answer correctly; examinees who are missing one or more skills are not differentiated from those who have no skills. Thus, each item partitions examinees into two latent groups, from which the slip (s) and guessing (g) parameters are estimated. The g parameter denotes the probability that an examinee who is missing one or more required skills provides a correct response; similarly, the s parameter denotes the probability that an examinee who possesses all required skills answers the question incorrectly. Occasionally, $1 - s$ is used rather than s; with $1 - s$, the interpretation becomes the probability that examinees who have all required skills answer the question correctly. Additional details for the DINA model can be found in de la Torre (2009b) and de la Torre and Douglas (2004).

Related to the DINA is the deterministic inputs, noisy "OR" gate (DINO; Templen & Henson, 2006) model. DINO items also partition examinees into two groups, but the two groups are defined differently than they are in the DINA model. The model only requires that examinees possess one required skill to provide a correct response; possessing additional skills does not change the probability of responding correctly. Only examinees who possess none of the required attributes are expected to respond incorrectly.

The generalized-DINA (G-DINA; de la Torre, 2011) model relaxes the simplifying constraints in the DINA and DINO models and allows for additional latent groups with intermediate probabilities of success to exist. Rather than partitioning examinees into two latent groups on each item as is done in the DINA and DINO models, the G-DINA partitions examinees into $2^{K_j^*}$ groups on each item j, where $K_j^* = \sum_{k=1}^{K} q_{jk}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, and J is the test length. For example, if $\mathbf{q}_j = 11010$, the G-DINA partitions examinees into eight groups for which unique probabilities of success can be estimated.

Although CDMs such as these represent important contributions to the literature, most CDMs can only handle dichotomous, and in a few cases, polytomous data (e.g., de la Torre, 2009a; Ma & de la Torre, 2016). To extract information with diagnostic value from the research settings involving continuous measures, a DINA-like CDM for continuous response will be introduced.

2. The DINA Model for Continuous Response

Let X_{ij} denote the response of examinee i to item j, where $X_{ij} \geq 0$, and $i = 1, \ldots, I$. In addition, let $\boldsymbol{\alpha}_i$ be the i^{th} attribute pattern, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$. Values of 1 and 0 on the k^{th} element of $\boldsymbol{\alpha}$ denote the presence and absence of attribute k, respectively. The model described below requires a Q-matrix (Tatsuoka, 1983), which is a binary $J \times K$ matrix where q_{jk}, the element in the j^{th} row and k^{th} column, indicates whether attribute k is required to correctly answer item j. For each examinee-item combination, a latent response variable η_{ij} is generated, which divides the examinees into two groups: η_{ij} assumes a value of 1 if examinee i possesses all the required attributes for item j, and 0 if the examinee lacks at least one of the required attributes. For convenience, η_{ij} will be denoted simply as η unless otherwise specified. The probability that response X_{ij} will be less than or equal to x, given the attribute pattern $\boldsymbol{\alpha}_i$ can be written as

$$P(X_{ij} \leq x|\boldsymbol{\alpha}_i) = \int_0^x [f_{j0}(x_{ij})]^{1-\eta_{ij}} [f_{j1}(x_{ij})]^{\eta_{ij}} dx_{ij}, \tag{1}$$

where

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{2}$$

and $f_{j\eta}(x_{ij})$ is the lognormal distribution, indexed by η, whose density is given as

$$f_{j\eta}(x_{ij}) = \frac{1}{x_{ij}\sqrt{2\pi\sigma_{j\eta}^2}} \exp\left[-\frac{(\log\ x_{ij} - \mu_{j\eta})^2}{2\sigma_{j\eta}^2}\right], \tag{3}$$

where $\mu_{j\eta} = E_{j\eta}(\log\ X_{ij})$ and $\sigma^2_{j\eta} = \mathrm{Var}_{j\eta}(\log\ X_{ij})$. That is, for each item, separate functions define the distributions of the responses for each group of η. Hereafter, the model will be referred to as the continuous DINA (C-DINA) model.

Note that the density given in Equation 3 may be replaced with any probability density function. For applications involving response time, such as those discussed in this article, a lognormal distribution has been shown to be appropriate (van der Linden, 2006). For applications in which the response is bounded (e.g., placing a mark on a line, probability testing, etc.), a distribution with finite support may be desired, such as the 2- or 4-parameter β distribution. Such an adjustment would represent a relatively minor interpretive extension to this model; however, substantial work would need to be done to revise the marginalized maximum likelihood estimation (MMLE; Bock & Aitkin, 1981) and standard error (*SE*) algorithms that are presented in Appendix A, particularly if the β distribution is desired due to the complexity of its first and second derivatives. To make this adjustment, the following changes would need to be made in addition to adjusting the notation to ensure its appropriateness: The derivatives in Equation 3 would need to be taken based on the distribution being used, and Equations 18, 19, 20, and 21 would need to be updated accordingly. It should also be noted that closed-form estimators may not be available depending on the distribution used. Updating the *SE* computation would involve changes to Equations 25 and 26. The simpler but more computationally intensive Markov chain Monte Carlo estimation technique could be used as well.

Three items with varying degrees of discrimination are shown in Figure 1. Items 1, 2, and 3 represent items with low, moderate, and high discriminations, respectively. The amount of overlap between the distributions corresponds inversely to the item's discrimination. An item's discrimination refers to its ability to differentiate the responses of examinees in $\eta = 0$ from those in $\eta = 1$ and thus can be viewed as a measure of item quality. Items with higher discrimination are expected to contribute to higher attribute classification accuracy. For illustration purposes, the responses in Figure 1 can be viewed as response time, where longer response time is an indication that examinees are successfully applying the skills required to solve the problem and thus is deemed better. That is, examinees who possess the required attributes for an item (i.e., those in $\eta = 1$) are expected to respond more slowly. Therefore, the density functions of the responses of examinees in $\eta = 0$ and $\eta = 1$ can be represented by the left and right curves of the plots, respectively. The parameters are given in Table 1.

To measure the discrimination of these items, an index that quantifies the difference between the two distributions is needed. A crude way to accomplish this is to compare the means of the two distributions, as in $\mu_1 - \mu_0$. This is analogous to computing $P(X = 1|\eta = 1) - P(X = 1|\eta = 0) = 1 - s - g$, an index that has been used to measure an item's discrimination in the context of the DINA model (de la Torre, 2008). However, it should be noted that when the

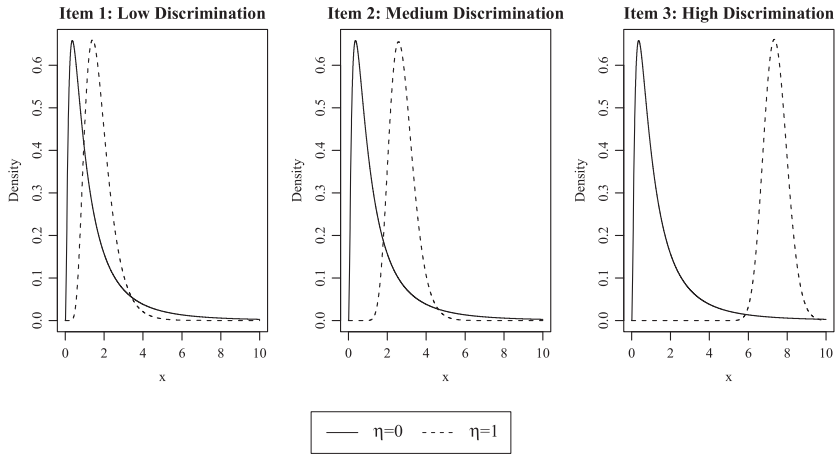

FIGURE 1. *Probability density curves of 3 items with different discriminations.*

TABLE 1.
Example Item Parameters

Item	Discrimination	μ_0	σ_0	μ_1	σ_1
1	Low	0	1	0.5	.397
2	Medium	0	1	1.0	.230
3	High	0	1	2.0	.082

variances of distributions vary substantially within and across items, this index may not accurately capture the discrimination power of an item, and a more involved index that takes into account the characteristics of the entire distributions is needed. One index for comparing similarity or dissimilarity of two distributions is the Kullback–Leibler index (KL; Cover & Thomas, 1991, as cited in Cheng, 2009). In the example items above, $\mu_1 - \mu_0$ for items 1, 2, and 3 are 0.5, 1, and 2, whereas the corresponding KLs are 24.15, 166.52, and infinite, respectively. Both indices indicate that Item 1 is the least discriminating and Item 3 is the most discriminating.

3. Estimation

The parameters of the C-DINA model, $\mu_{j\eta}$ and $\sigma^2_{j\eta}$, for $\eta = 0$ and 1, $j = 1, \ldots, J$, can be estimated using MMLE. As shown in Appendix A, the estimators of these parameters are

$$\hat{\mu}_{j\eta} = \sum_{i=1}^{I} p_{ij}(\eta) \log x_{ij}, \tag{4}$$

and

$$\hat{\sigma}^2_{j\eta} = \sum_{i=1}^{I} p_{ij}(\eta)(\log x_{ij} - \hat{\mu}_{j\eta})^2, \tag{5}$$

where $p_{ij}(\eta) = p(\eta_j = \eta|\mathbf{x}_i)/\sum_{i=1}^{I} p(\eta_j = \eta|\mathbf{x}_i)$ and $p(\eta_j = \eta|\mathbf{x}_i)$ is the posterior probability that examinee i is in group η with respect to item j.

Similarly, the *SE*s of the parameter estimates of item j can be approximated by the square root of the diagonal elements of

$$\left(\sum_{i=1}^{I} \frac{p^2(\eta_j = \eta|\mathbf{x}_i)}{4(\sigma^2_{j\eta})^2} \mathbf{d}_{j\eta} \mathbf{d}'_{j\eta} \right)^{-1}, \tag{6}$$

where

$$\mathbf{d}_{j\eta} = \begin{pmatrix} 2(\log x_{ij} - \mu_{j\eta}) \\ (\log x_{ij} - \mu_{j\eta})^2/\sigma^2_{j\eta} - 1 \end{pmatrix}. \tag{7}$$

For the approximation implemented in this article, Equation 6 involves inversion of a diagonal matrix.

4. Simulation Study

A simulation study was carried out to investigate the viability of the proposed C-DINA model. Specifically, the study was designed to determine how estimation of the C-DINA model parameters and its attribute classification rate are affected by three factors—test length (J), sample size (I), and item discrimination. Additionally, the study examined how the attribute classification accuracy of the C-DINA model compared to that of the DINA model when the continuous response was dichotomized. The code to estimate the model was written in R (R Core Team, 2015).

4.1. Design and Analysis

Two test lengths ($J = 15, 30$), three sample sizes ($I = 500, 1{,}000, 2{,}000$), and the three discriminations presented in Table 1 were considered. Note that only the parameters for $\eta = 1$ were varied across the conditions, whereas the parameters for $\eta = 0$ were fixed at $\mu_0 = 0.00$ and $\sigma^2_0 = 1.00$. For greater comparability, all the items in the test were assigned identical parameters for a given level of discrimination. The Q-matrices used in the simulation study are given in Table 2, with * denoting items that were used in the 15-item test. For each condition, 100 data sets were simulated, and model parameters were estimated using the previously discussed EM procedure. The convergence criterion, which was defined as the maximum absolute difference between the current and

TABLE 2.
Simulation Study Q-matrix

	Attribute						Attribute				
Item	α_1	α_2	α_3	α_4	α_5	Item	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

Note. Asterisk (*) denotes items used in the $J = 15$ conditions.

previous parameter estimates across all items, was fixed at 0.0001. All replications converged for all conditions.

The quality of the item parameter estimates for a given combination of conditions was summarized by computing the mean bias and variability across the 100 replications. Using the marginal probabilities of the posterior distributions to classify the examinees, classification accuracy was computed at both the attribute and vector levels. For comparison purposes, the generated responses were reduced to dichotomous data and analyzed using the DINA model. The objective of the dichotomization rule was to maximize the separation between the response curve densities of groups $\eta = 0$ and 1; thus, the point of dichotomization was the point at which the two densities intersected. For Items 1 through 3, in Figure 1, the dichotomization points along the response axis that produce the optimal separation are 0.96, 1.79, and 5.85, respectively. Responses to the left of the cutoff points were scored as 0, and 1 otherwise. Reducing the data in this way provided the maximum distinction between the 0s and 1s, which made it more similar to dichotomous data than any other method of dichotomization.

4.2. Results

Tables 3 and 4 summarize the results for the estimates of μ_η, whereas Tables 5 and 6 summarize the results for the estimates of σ^2_η, for $\eta = 0, 1$, respectively.

TABLE 3.
Mean, Bias, and SE *and* SD *of* $\hat{\mu}_0$ *Over 100 Replications*

			Number of Required Attributes								
			One			Two			Three		
J	(μ_1,σ_1)	I	Bias	*SE*	*SD*	Bias	*SE*	*SD*	Bias	*SE*	*SD*
		500	−.01	.07	.08	−.01	.05	.06	.00	.05	.05
	(0.50, 0.40)	1,000	.00	.05	.06	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.04	.00	.03	.03	.00	.02	.02
		500	.00	.07	.07	.00	.05	.05	.00	.05	.05
15	(1.00, 0.23)	1,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.04	.00	.03	.03	.00	.02	.02
		500	.00	.06	.06	−.01	.05	.05	.00	.05	.05
	(2.00, 0.08)	1,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.03	.00	.03	.03	.00	.02	.02
		500	.00	.07	.07	.00	.05	.05	.00	.05	.05
	(0.50, 0.40)	1,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.03	.00	.03	.03	.00	.02	.02
		500	.00	.06	.06	.00	.05	.05	.00	.05	.05
30	(1.00, 0.23)	1,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.03	.00	.03	.03	.00	.02	.02
		500	.00	.06	.06	.00	.05	.05	.00	.05	.05
	(2.00, 0.08)	1,000	.00	.04	.04	.00	.04	.04	.00	.03	.03
		2,000	.00	.03	.03	.00	.03	.03	.00	.02	.02

Note. *SE* = standard error; *SD* = standard deviation.

The results presented for each condition are the mean of the biases, the square root of the mean of the analytical *SE*s (computed using Equation 6), and the square root of the mean of the empirical variances. These quantities are referred to in the tables as bias, *SE*, and standard deviation, respectively. Means were taken across the 100 replications and across items that require the same number of attributes. For example, the results in columns 4 through 6 pertain to Items 1 through 5 or 1 through 10 for the $J = 15$ and $J = 30$ conditions, respectively, which represented items requiring only a single attribute, in accordance with Table 2.

Virtually all conditions resulted in unbiased estimates, with just a few conditions resulting in a small negative bias of 0.01. The analytical *SE*s that were computed using Equation 6 were very close to the corresponding empirical *SE*s of the estimates. At most, they differed by less than 0.02 but frequently were identical to the second decimal place. Thus, we will discuss variability in general without making a distinction between these two measures. There was also a consistent pattern to the variability with respect to the number of attributes and η.

TABLE 4.
Mean, Bias, and SE *and* SD *of* $\hat{\mu}_1$ *Over 100 Replications*

			Number of Required Attributes								
			One			Two			Three		
J	(μ_1,σ_1)	I	Bias	*SE*	*SD*	Bias	*SE*	*SD*	Bias	*SE*	*SD*
		500	.00	.03	.03	.00	.05	.05	.00	.07	.08
	(0.50, 0.40)	1,000	.00	.02	.02	.00	.03	.03	.00	.05	.05
		2,000	.00	.02	.02	.00	.02	.03	.00	.03	.03
		500	.00	.02	.02	.00	.02	.02	.00	.03	.03
15	(1.00, 0.23)	1,000	.00	.01	.01	.00	.02	.02	.00	.02	.02
		2,000	.00	.01	.01	.00	.01	.01	.00	.02	.02
		500	.00	.01	.01	.00	.01	.01	.00	.01	.01
	(2.00, 0.08)	1,000	.00	.00	.00	.00	.01	.01	.00	.01	.01
		2,000	.00	.00	.00	.00	.00	.00	.00	.01	.01
		500	.00	.03	.03	.00	.04	.04	.00	.06	.06
	(0.50, 0.40)	1,000	.00	.02	.02	.00	.03	.03	.00	.04	.04
		2,000	.00	.01	.01	.00	.02	.02	.00	.03	.03
		500	.00	.01	.02	.00	.02	.02	.00	.03	.03
30	(1.00, 0.23)	1,000	.00	.01	.01	.00	.01	.01	.00	.02	.02
		2,000	.00	.01	.01	.00	.01	.01	.00	.01	.01
		500	.00	.01	.01	.00	.01	.01	.00	.01	.01
	(2.00, 0.08)	1,000	.00	.00	.00	.00	.01	.01	.00	.01	.01
		2,000	.00	.00	.00	.00	.00	.00	.00	.01	.01

Note. *SE* = standard error; *SD* = standard deviation.

For μ_0 and σ_0, variability decreased as the number of attributes increased, whereas for μ_1 and σ_1, variability increased as the number of attributes, K_j^*, increased. This is due to the number of attribute patterns comprising the $\eta = 0$ and $\eta = 1$ groups—as the number of attributes measured by the item increases, the number of attribute patterns in the $\eta = 0$ group increases whereas it decreases for the $\eta = 1$ group. For example, Item 1 splits the patterns equally, where there are 16 patterns in each group. Item 11, however, splits the patterns into 8 in the $\eta = 1$ group and the remaining 24 in the $\eta = 0$ group. Under Item 21, there are just 4 patterns in $\eta = 1$ whereas there are 28 in $\eta = 0$. In general, the number of attribute patterns in the $\eta = 0$ and $\eta = 1$ groups for DINA models is $2^K - 2^K/2^{K_j^*}$ and $2^K/2^{K_j^*}$, respectively. A greater number of patterns in one of the groups results in a more stable (i.e., less variable) estimation of the relevant parameter. In addition to this global pattern, other more specific patterns were observed.

Table 3 indicates that variability in $\hat{\mu}_0$ decreased as either the sample size or discrimination was increased, but increasing the sample size had a more dramatic effect. For a given discrimination and attribute condition, increasing the sample

TABLE 5.
Mean, Bias, and SE *and* SD *of* $\hat{\sigma}_0^2$ *Over 100 Replications*

			Number of Required Attributes								
			One			Two			Three		
J	(μ_1, σ_1)	*I*	Bias	*SE*	*SD*	Bias	*SE*	*SD*	Bias	*SE*	*SD*
		500	.00	.10	.11	.00	.08	.07	.00	.07	.07
	(0.50, 0.40)	1,000	.00	.07	.08	.00	.05	.06	.00	.05	.05
		2,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		500	−.01	.09	.09	.00	.07	.07	.00	.07	.07
15	(1.00, 0.23)	1,000	−.01	.06	.07	.00	.05	.05	.00	.05	.05
		2,000	.00	.05	.04	.00	.04	.04	.00	.03	.03
		500	−.01	.09	.09	.00	.07	.07	.00	.07	.07
	(2.00, 0.08)	1,000	.00	.07	.06	.00	.05	.05	.00	.05	.05
		2,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		500	−.01	.09	.09	.00	.08	.07	.00	.07	.07
	(0.50, 0.40)	1,000	.00	.07	.06	.00	.05	.05	.00	.05	.05
		2,000	.00	.05	.05	.00	.04	.04	.00	.03	.03
		500	−.01	.09	.09	.00	.07	.07	.00	.07	.07
30	(1.00, 0.23)	1,000	.00	.06	.06	.00	.05	.05	.00	.05	.05
		2,000	.00	.05	.04	.00	.04	.04	.00	.03	.03
		500	.00	.09	.09	.00	.07	.08	.00	.07	.07
	(2.00, 0.08)	1,000	.00	.06	.06	.00	.05	.05	.00	.05	.05
		2,000	.00	.05	.05	.00	.04	.04	.00	.03	.03

Note. *SE* = standard error; *SD* = standard deviation.

size reduced variability by as much as 0.04, while for a given sample size and attribute condition, increasing the discrimination only resulted in a reduction of at most 0.02, and this only occurred for one-attribute items. Test length did not have an effect on variability.

Compared with $\hat{\mu}_0$, the variability of the $\hat{\mu}_1$, shown in Table 4, was more related to simulation conditions. Increasing the sample size or increasing the discrimination both resulted in reductions in the variability of the estimate. However, increasing the discrimination reduced the variability by as much as 0.07, which was almost twice the magnitude of the reduction for $\hat{\mu}_1$. Increasing the sample size resulted in a reduction of variability by as much as 0.05. The most variable conditions were those with smaller sample size, lower discrimination, and more attributes. Increasing the test length also appeared to reduce variability somewhat, but this effect was most apparent for the low discrimination conditions. Maximum reductions were only 0.02.

Similar to the other parameters, Table 5 indicates that variability for σ_0^2 decreased when the sample size and discrimination were increased, with the

TABLE 6.
Mean, Bias, and SE *and* SD *of* $\hat{\sigma}_1^2$ *Over 100 Replications*

			Number of Required Attributes								
			One			Two			Three		
J	(μ_1,σ_1)	*I*	Bias	*SE*	*SD*	Bias	*SE*	*SD*	Bias	*SE*	*SD*
		500	.00	.02	.03	.00	.03	.04	.01	.05	.06
	(0.50, 0.40)	1,000	.00	.01	.02	.00	.02	.02	.00	.03	.04
		2,000	.00	.01	.01	.00	.02	.02	.00	.02	.02
		500	.00	.01	.01	.00	.01	.01	.00	.01	.01
15	(1.00, 0.23)	1,000	.00	.00	.00	.00	.01	.01	.00	.01	.01
		2,000	.00	.00	.00	.00	.00	.00	.00	.01	.01
		500	.00	.00	.00	.00	.00	.00	.00	.00	.00
	(2.00, 0.08)	1,000	.00	.00	.00	.00	.00	.00	.00	.00	.00
		2,000	.00	.00	.00	.00	.00	.00	.00	.00	.00
		500	.00	.02	.02	.00	.03	.03	.00	.04	.03
	(0.50, 0.40)	1,000	.00	.01	.01	.00	.02	.02	.00	.02	.02
		2,000	.00	.01	.01	.00	.01	.01	.00	.02	.02
		500	.00	.01	.00	.00	.01	.01	.00	.01	.01
30	(1.00, 0.23)	1,000	.00	.00	.00	.00	.01	.00	.00	.01	.01
		2,000	.00	.00	.00	.00	.00	.00	.00	.00	.00
		500	.00	.00	.00	.00	.00	.00	.00	.00	.00
	(2.00, 0.08)	1,000	.00	.00	.00	.00	.00	.00	.00	.00	.00
		2,000	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note. *SE* = standard error; *SD* = standard deviation.

more noticeable reductions occurring as sample size increased. Increasing the discrimination did not have an effect for three-attribute items; it did, however, reduce the variance slightly in some conditions for one- and two-attribute items by at most 0.02. Increasing the test length generally did not have a noticeable effect on variability.

The results for $\hat{\sigma}_1^2$ are shown in Table 6. Noticeable variability was only present for low discrimination conditions; for higher discriminations, variability was 0.01 or less. For the low discrimination condition, increasing the sample size reduced variability by as much as 0.04, and increasing the test length reduced variability by as much as 0.03.

Finally, the correct attribute- and vector-wise classification rates are shown in Table 7. In contrast to its effect on parameter estimation, sample size had a minimal (but positive) impact on classification whereas discrimination played a much more prominent role—increasing the separation between the responses of the two groups dramatically increased the correct classification rates for both models. Within a test length, increasing discrimination increased the

TABLE 7.
Mean Classification Accuracy Rates for the DINA and C-DINA Models

			Classification Type					
			Single			Vector		
J	(μ_1,σ_1)	I	C-DINA	DINA	Gain	C-DINA	DINA	Gain
		500	77.25	69.14	8.11	35.85	20.55	15.30
	(0.50, 0.40)	1,000	78.12	70.20	7.92	38.16	22.40	15.76
		2,000	78.60	70.79	7.81	39.30	23.68	15.62
		500	91.99	87.43	4.56	72.86	59.40	13.46
15	(1.00, 0.23)	1,000	92.22	87.85	4.37	73.53	60.66	12.87
		2,000	92.37	88.13	4.24	73.88	61.62	12.26
		500	99.39	98.88	0.51	97.54	95.72	1.82
	(2.00, 0.08)	1,000	99.44	98.87	0.57	97.64	95.72	1.93
		2,000	99.40	98.85	0.56	97.54	95.62	1.92
		500	89.30	79.92	9.39	65.64	40.77	24.87
	(0.50, 0.40)	1,000	89.68	80.68	9.00	66.97	42.97	24.00
		2,000	89.77	80.96	8.81	67.25	43.72	23.53
		500	98.54	95.79	2.74	94.48	84.43	10.05
30	(1.00, 0.23)	1,000	98.57	95.96	2.60	94.63	85.06	9.57
		2,000	98.57	96.02	2.56	94.70	85.24	9.46
		500	99.99	99.92	0.08	99.97	99.60	0.37
	(2.00, 0.08)	1,000	99.99	99.92	0.07	99.97	99.59	0.37
		2,000	99.99	99.92	0.07	99.97	99.63	0.34

Note. DINA = deterministic inputs, noisy "AND" gate; C-DINA = continuous deterministic inputs, noisy "AND" gate.

classification accuracy for both models, but at a slower rate for the C-DINA, meaning that the improvement over the DINA model lessened. Generally, increasing the test length resulted in improved classification, but doing so affected the models differently, resulting in different levels of gain. For example, for the low discrimination condition, increasing the test length improved classifications, but did so to a greater degree for the C-DINA, resulting in an increased gain. For medium and high discrimination conditions, increasing the test length improved accuracy for the DINA model more than for the C-DINA, resulting in slightly lower gains for the longer test. This could be due to the fact that the C-DINA rapidly approached classification accuracy in the mid-to high-90 percentage range as the favorability of the conditions improved. The gain offered by the C-DINA model diminished to less than 2 percentage points at the vector level and less than 1 percentage point for the attribute level when the discrimination was high.

In comparison to results obtained using dichotomized data, the classification rates using continuous data were always better. The differences were greatest when the discrimination was low and when considering vector-level

classification accuracy. As the items became more discriminating, the differences between the DINA and C-DINA models became negligible.

5. Real Data Example

5.1. Data Description

Van der Maas and Jansen (2003) suggested that examining response times may provide additional insight into the cognitive processes that underlie a test beyond what response accuracy data shows. Siegler (1989) also suggested that response time analyses may be very useful in examining the strategies that examinees use to solve a problem. As such, we applied both the DINA and C-DINA models to a data set of responses and response times of students on balance scale tasks collected and originally analyzed by van der Maas and Jansen (2003).

On the balance scale task, a participant is asked to predict the movement of the scale—tilting to the left, tilting to the right, or balanced. On both arms of the scale, pegs are situated at equal distances from each other and from the fulcrum. On each side of the fulcrum, one or more identical weights are placed on a single peg; the weights placed on each side need not be equal in number nor in distance from the fulcrum. In fact, it is the differences in the numbers and the distances that define the problem type. Participants' responses to the various types of balance scale items reveal the sets of skills they possess.

The balance scale problems require students to employ increasingly complex skills, which can be viewed as being developmental in nature (Siegler, 1976, 1981). Van der Maas and Jansen (2003) derived the required steps for each item type from the basic model proposed by Siegler (1981). The original data collected by van der Maas and Jansen (2003) consisted of responses from 191 students (147 primary and secondary school and 44 college students) on eight item types, for which the data were complete on seven. To apply the models, we analyzed the responses and response times (in seconds) of the primary and secondary students on four of the seven item types. We removed examinees with ages greater than 16, and those for whom no age was recorded, resulting in a final sample of 146 students.

We began by using all seven item types, of which three were removed. One was removed because of a relationship with age that was opposite to that of the other question types. Two more were removed because they were very easy with proportions correct in excess of 0.9 (van der Maas & Jansen, 2003, p. 156); generally, these item types do not differentiate well between the $\eta = 0$ and $\eta = 1$ groups.

The Q-matrix was constructed based on the following steps required to solve each problem type: (1) comparing the *distances* at which the weights are placed and (2) applying the *torque* rule (van der Maas & Jansen, 2003). In the context of CDM, these steps may be interpreted as attributes. Van der Maas and Jansen (2003) hypothesized that the time required to complete a question is directly

TABLE 8.
Q-matrix for the Balance Scale Data

		Attribute	
Item Type	Description	Distance	Torque
I	Simple distance	1	0
II	Conflict balance B	1	1
III	Conflict distance	1	1
IV	Conflict balance A	1	1

related to the steps required to solve it correctly. Increasingly, complex items require the application of complex steps beyond the required basic steps and thus require more time to complete. Therefore, the Q-matrix used for response accuracy analyses can also be used response time analyses.

The resulting Q-matrix is presented in Table 8. There were 10 items for each of the four item types for a total of 40 items. Because this application involved developmental attributes, we constrained the possible attribute patterns such that mastery of an attribute implies mastery of all lower level attributes. Thus, in this study, only three such patterns were estimated: $\boldsymbol{\alpha} = 00$, $\boldsymbol{\alpha} = 10$, and $\boldsymbol{\alpha} = 11$. As in the simulation study, the convergence criterion was defined as 0.0001 for both models, and the data converged after 16 and 29 iterations for the DINA and C-DINA models, respectively.

Although our primary goal was to analyze response time using the C-DINA model, we began by applying the DINA model to the response accuracy to determine the starting values for the C-DINA EM algorithm. With the DINA classifications, we partitioned the examinees into $\eta = 0$ and 1 groups according to the item type. We then examined the distribution of each examinee's response time averaged within a given item type. Thus, there were eight distributions examined: those for the $\eta = 0$ and 1 groups for each of the four item types. The means and the variances of the distributions of log times were used as the starting values for the C-DINA algorithm.

5.2. Results

The C-DINA model parameter estimates and the corresponding *SE*s for these data can be found in Table 9. Instead of presenting the estimates for the 40 items individually, the table lists the mean estimates across the 10 items for each item type. Unlike the normal distribution, the first and second moments of the lognormal distribution are dependent upon both μ and σ parameters. Specifically, $E(X) = (e^{\mu+\sigma^2})$ and $\text{Var}(X) = (e^{\sigma^2} - 1)(e^{2\mu+\sigma^2})$ for lognormal X. Thus, when both $\hat{\mu}_a > \hat{\mu}_b$ and $\hat{\sigma}_a^2 > \hat{\sigma}_b^2$, both the mean and the variance of X for $\eta = a$ are

TABLE 9.
C-DINA Model Parameter Estimates for the Balance Scale Data

	Parameter			
Item Type	μ_0	σ_0^2	μ_1	σ_1^2
I	1.00 (0.05)	.12 (.03)	1.41 (0.07)	.26 (.05)
II	1.06 (0.05)	.16 (.02)	1.76 (0.09)	.34 (.08)
III	1.13 (0.05)	.16 (.03)	1.72 (0.08)	.33 (.07)
IV	1.02 (0.05)	.18 (.03)	1.72 (0.09)	.38 (.09)

Note. Standard errors are in parentheses. C-DINA = continuous deterministic inputs, noisy "AND" gate.

larger than for $\eta = b$. In this example, the μ_η and σ_η estimates show that the mean response times of students in $\eta = 1$ were both longer and more variable for all item types compared to those of students in $\eta = 0$. While the longer mean times were expected based on theory, the increased variability may have been natural result of the lack of an upper bound. Although not shown here, these findings were also true for all 40 items individually.

Figure 2 shows average response times by item type for each of the latent groups and classes; the plots on the left classified examinees according to their response accuracy analyzed by the DINA model, whereas the plots on the right classified examinees according to their response time analyzed by the C-DINA model. As discussed earlier, both the DINA and C-DINA models partition examinees into just two latent groups—$\eta = 0$ and 1—on each item. Response times analyzed at this level are shown in the top of the figures. In the bottom of the figure, response times are shown for each of the three estimated latent classes. By examining the latent groups first (top), we can see that the general pattern is very similar regardless of the CDM used. The separation in the response times was greater for the $\mathbf{q}_j = 11$ problems than it was for the $\mathbf{q}_j = 10$, as expected, due to the fact that the $\mathbf{q}_j = 11$ problems require a second skill to be used. Next, the latent classes (bottom) revealed a more detailed pattern, which was also similar for each of the models. In both plots, the $\boldsymbol{\alpha} = 10$ performed similarly to the $\boldsymbol{\alpha} = 11$ group for the $\mathbf{q}_j = 10$ problems, which was expected because they are both in $\eta = 1$. For the $\mathbf{q}_j = 11$ problems, the $\boldsymbol{\alpha} = 10$ group performed similarly to the $\boldsymbol{\alpha} = 00$ because they are both in $\eta = 0$. The main difference between the DINA plots on the left and the C-DINA plots on the right is that the response times in the C-DINA are slightly further separated than they are for the DINA, which was expected because the C-DINA maximizes the differences in response times.

The format of Figure 3 is similar to that of Figure 2, except that it shows average response accuracy (proportion correct) rather than response time by item type for each of the latent groups and classes. Examining the top two plots, which

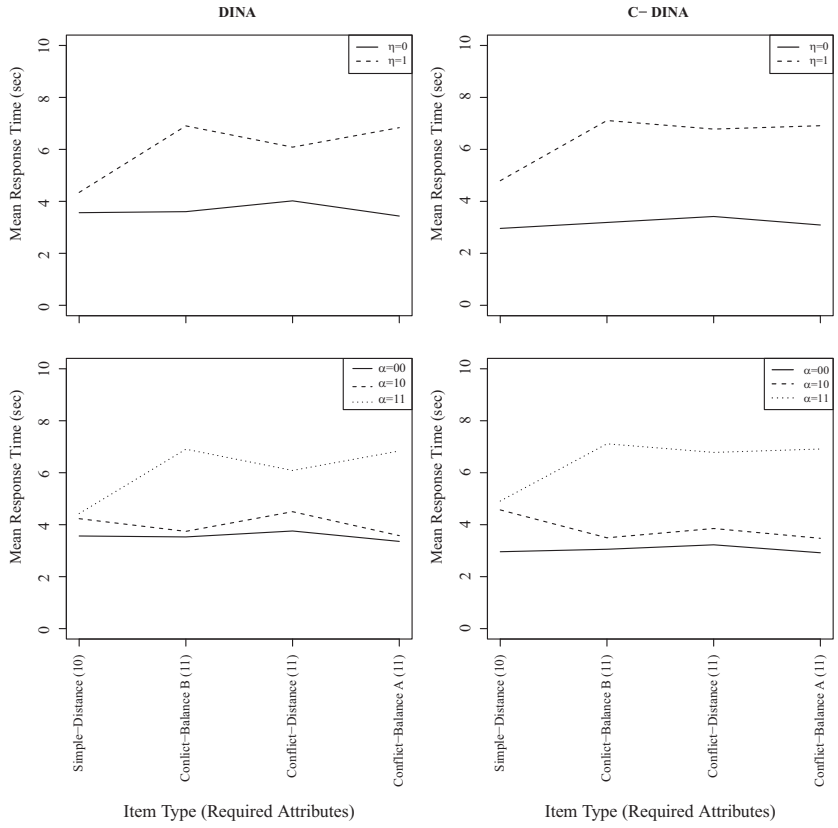

FIGURE 2. *The deterministic inputs, noisy "AND" gate (DINA, left) and C-DINA (right) mean response times by latent group (upper) and class (lower).*

display the proportion correct as a function of the latent groups, shows that the patterns are again quite similar, particularly for $\eta = 1$. The most notable difference is that the $\eta = 0$ group had a proportion correct of approximately 0.3 for the $\mathbf{q}_j = 10$ items when the C-DINA classification was used, whereas the corresponding proportion correct when the DINA classification was used was just slightly above 0.

In examining the figure on the bottom left, the $\boldsymbol{\alpha} = 10$ group's proportion correct was nearly identical to either $\boldsymbol{\alpha} = 11$ (first item type) or $\boldsymbol{\alpha} = 00$ (last three item types) depending on whether $\boldsymbol{\alpha} = 10$ was in $\eta = 0$ or $\eta = 1$. These similarities were expected because the DINA model maximizes the differences in proportions correct between the $\eta = 0$ and 1 groups. The results for the first item type were different across the models. For the DINA classification, the

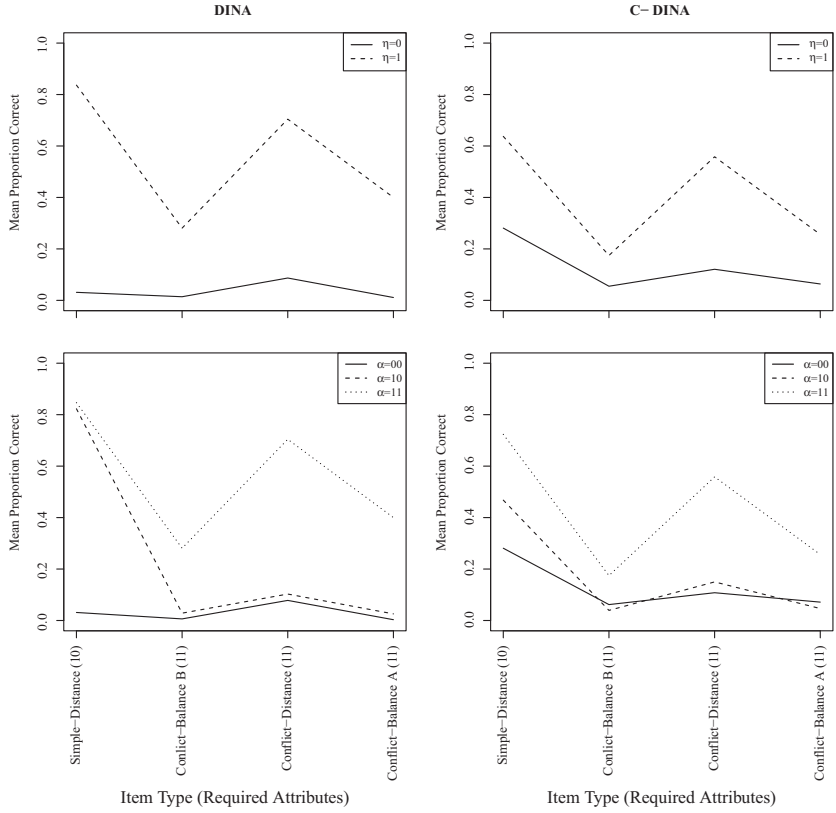

FIGURE 3. *The deterministic inputs, noisy "AND" gate (DINA, left) and C-DINA (right) mean proportions correct by latent group (upper) and class (lower).*

$\boldsymbol{\alpha} = 10$ and $\boldsymbol{\alpha} = 11$ groups had accuracies of around 0.8, whereas the $\boldsymbol{\alpha} = 00$ group's accuracy was about 0.05. Using C-DINA classifications, the accuracies of the three latent classes are roughly equidistant, and not as extreme. These latent classes were, however, ordered such that examinees with more attributes were more likely to answer correctly.

Lastly, Table 10 shows the cross tabulation of the classifications for the two models. There were 92 students (63%) who had the same classifications for both models. Classifications that differed by only one attribute were observed for 37 students (25%), and the remaining 17 students (12%) had classifications that differed on both attributes. These data provide rich information into the response processes for students.

Let $\boldsymbol{\alpha}_{\text{RT}}$ and $\boldsymbol{\alpha}_{\text{RA}}$ denote students' C-DINA and DINA classifications, respectively. The upper triangle of the matrix represents students whose $\boldsymbol{\alpha}_{\text{RA}}$ had fewer

TABLE 10.
Confusion Matrix for DINA and C-DINA Classifications

	C-DINA (RT)			
DINA (RA)	00	10	11	Total
00	43	12	9	64
10	12	13	10	35
11	8	3	36	47
Total	63	28	55	146

Note. DINA = deterministic inputs, noisy "AND" gate; C-DINA = continuous deterministic inputs, noisy "AND" gate; RA = response accuracy; RT = response time.

attributes than their $\boldsymbol{\alpha}_{RT}$. Their response time profiles suggested they were applying one or both attributes but were not doing so successfully, an observation that could lead to various hypotheses to explain the discrepancy. Conversely, the lower triangle of the matrix represents students whose $\boldsymbol{\alpha}_{RT}$ had fewer attributes than their $\boldsymbol{\alpha}_{RA}$. These were students who appeared to be able to execute either one or both of the rules very quickly and thus correctly responded to certain problems more quickly than other students who also responded correctly to similar problems.

6. Discussion and Conclusion

As computer-based testing becomes more prevalent, continuous response data will continue to become more readily available. This article proposed the C-DINA model, a CDM for continuous response. With the C-DINA model, latent variable modeling of continuous response data need not be based on IRT models that rely on a single continuous trait, which can only indicate a respondent's overall standing. As an alternative, these data can be analyzed using models that can provide more specific information regarding the respondent's standing in a multidimensional space. As with CDMs for discrete data, the C-DINA model has the potential to provide inferences that are richer and more informative and thus more useful from both theoretical and practical perspectives.

The simulation study supported the viability of the proposed model. Using the computer code based on an implementation of the EM algorithm, the study showed that reasonable estimates of the C-DINA model parameters can be obtained even with a relatively small sample size. A larger sample size is recommended if parameter estimates of greater precision are desired or if tests with less discriminating items are involved. All four parameters were affected by sample size and discrimination, whereas test length primarily affected $\eta = 1$ parameters. Sample size tended to have a larger effect for $\eta = 0$ parameters, whereas discrimination had a larger effect for $\eta = 1$ parameters. The simulation

study also showed that the correct attribute classification rates based on continuous responses were always better than those based on dichotomous responses, particularly at the vector level and with items of lower discrimination.

Applying the DINA and C-DINA models to the response times on the balance scale tasks revealed several interesting findings. By applying the expectation and variance formulas for the lognormal distribution, the estimated model parameters revealed both the mean and the variability of the response times across the different item types, allowing for inferences to be made about the characteristics of the responses of each group (i.e., $\eta = 0$ or $\eta = 1$). Additionally, the students' response time profiles indicated that students with different attribute patterns responded differentially to the different item types. The patterns of response time and response accuracy for the DINA and C-DINA models were quite similar overall, with only substantial differences on the response accuracy for one item type.

The cross tabulation of classification showed that the classifications for the models were related, although they were not identical. This provides a basis for using one response type to augment the other, although further study is needed. It can also be used to better understand why certain groups of examinees respond in a particular way. Some groups of examinees appeared to understand certain concepts based on their response time, but failed to implement them correctly. Conversely, other students appeared to solve complex item types very quickly. These observations underscore a practical matter when applying the C-DINA model: If educators are interested in response time but response accuracy is inherently valuable, a dual analysis of the response accuracy and the response time should be performed. Such an analysis provides richer information than a single-response analysis.

The work represented in this article is an initial attempt at better understanding continuous response from a cognitive diagnosis perspective. Although it broadens our understanding of the proposed model in particular and cognitive diagnosis modeling of continuous data in general, a lot of work remains to be done in this area.

First, the current setup of the simulation study presented in this article was limited in many ways. In particular, the number of attributes, the parameters of the items for a given discrimination condition, and the Q-matrix were all fixed in this study. In future studies, they can be considered as additional factors, and their impact on item parameters estimates and attribute classification accuracy in the continuous response context should be examined. Next, the assumption of the C-DINA model used to formulate Equation 2 can be relaxed, and its formulation can be generalized in a manner analogous to the way that the DINA model was generalized by de la Torre (2011). In such a model, each item partitions examinees into $2^{K_j^*}$ latent groups. Thus, $2^{K_j^*}$ lognormal distributions would be estimated per item, requiring the estimation of $2 \cdot 2^{K_j^*} = 2^{K_j^*+1}$ parameters rather than just $4J$.

Third, additional work needs to be done in examining the fit of the C-DINA model, particularly to real data. The inferences based on this model are only valid

to the extent to which it provides an adequate fit to the data. One way of examining model–data fit is to validate the Q-matrix specifications. For example, de la Torre (2008) developed an empirical Q-matrix validation method for the DINA model, and de la Torre and Chiu (2016) extended this method to the G-DINA (de la Torre, 2011) model. Similarly, Q-matrix validation methods could be developed for the standard or generalized version of the C-DINA model.

Finally, the current formulation of the C-DINA model limits its application strictly to continuous response. As alluded to earlier, the model can be extended to handle continuous and dichotomous data simultaneously. Such a model can build upon the work of van der Linden (2007), whose model from a hierarchical IRT framework can simultaneously handle both the latency and the correctness of the response.

Appendix A

Estimation of the C-DINA Model Parameters via an EM Algorithm

Estimating the item parameters via MMLE. The conditional likelihood of $\mathbf{x}_i$ given $\boldsymbol{\alpha}_i$ is

$$L(\mathbf{x}_i|\boldsymbol{\alpha}_i) = \prod_{j=1}^{J}[f_{j0}(x_{ij})]^{1-\eta_{ij}}[f_{j1}(x_{ij})]^{\eta_{ij}}, \tag{8}$$

and the corresponding marginalized likelihood is

$$L(\mathbf{x}_i) = \sum_{h=1}^{H} L(\mathbf{x}_i|\boldsymbol{\alpha}_h)p(\boldsymbol{\alpha}_h), \tag{9}$$

$H = 2^K$, and $p(\boldsymbol{\alpha}_h)$ is the prior density of α_h. Finally, the marginalized likelihood of the data $\mathbf{X}$ is

$$L(\mathbf{X}) = \prod_{i=1}^{I} L(\mathbf{x}_i). \tag{10}$$

Define $\boldsymbol{\Phi} = (\phi_1', \ldots, \phi_J')'$, where $\phi_j = (\phi_{j0}', \phi_{j1}')'$ and $\phi_{j\eta} = (\mu_{j\eta}, \sigma_{j\eta}^2)'$. To obtain $\widehat{\boldsymbol{\Phi}}$, the MMLE of $\boldsymbol{\Phi}$, we need to maximize Equation 10, or equivalently, the log-marginal likelihood

$$l(\mathbf{X}) = \log L(\mathbf{X}) = \sum_{i=1}^{I} \log L(\mathbf{x}_i), \tag{11}$$

with respect to $\boldsymbol{\Phi}$. The derivative of Equation 11 can be carried out separately by item and the subset of the parameters within an item. Specifically, the derivative of Equation 11 can be taken with respect to $\phi_{j\eta}$. That is,

$$\frac{\partial l(\mathbf{X})}{\partial \phi_{j\eta}} = \sum_{i=1}^{I} \frac{\partial L(\mathbf{x}_i)}{\partial \phi_{j\eta}} \Big/ L(\mathbf{x}_i) = \sum_{i=1}^{I} \frac{1}{L(\mathbf{x}_i)} \sum_{h=1}^{H} p(\alpha_h) \frac{\partial L(\mathbf{x}_i|\boldsymbol{\alpha}_h)}{\partial \phi_{j\eta}}. \tag{12}$$

The derivative of the conditional likelihood $L(\mathbf{x}_i|\boldsymbol{\alpha}_h)$ with respect to $\phi_{j\eta}$ is

$$\left\{ \prod_{j \neq j^*} [f_{j^*0}(x_{ij^*})]^{1-\eta_{hj^*}} [f_{j^*1}(x_{ij^*})]^{\eta_{hj^*}} \right\} \frac{\partial [f_{j0}(x_{ij})]^{1-\eta_{hj}} [f_{j1}(x_{ij})]^{\eta_{hj}}}{\partial \phi_{j\eta}}, \tag{13}$$

and can be shown to be equal to

$$L(\mathbf{x}_i|\boldsymbol{\alpha}_h) \left[\frac{1-\eta_{hj}}{f_{j0}(x_{ij})} \cdot \frac{\partial f_{j0}(x_{ij})}{\partial \phi_{j\eta}} + \frac{\eta_{hj}}{f_{j1}(x_{ij})} \cdot \frac{\partial f_{j1}(x_{ij})}{\partial \phi_{j\eta}} \right]. \tag{14}$$

Using Equation 14, Equation 12 can be written as

$$\begin{aligned} &\sum_{i=1}^{I} \frac{1}{L(\mathbf{x}_i)} \sum_{h=1}^{H} p(\boldsymbol{\alpha}_h) L(\mathbf{x}_i|\boldsymbol{\alpha}_h) \left[\frac{1-\eta_{hj}}{f_{j0}(x_{ij})} \cdot \frac{\partial f_{j0}(x_{ij})}{\partial \phi_{j\eta}} + \frac{\eta_{hj}}{f_{j1}(x_{ij})} \cdot \frac{\partial f_{j1}(x_{ij})}{\partial \phi_{j\eta}} \right] \\ &= \sum_{i=1}^{I} \sum_{h=1}^{H} \frac{L(\mathbf{x}_i|\boldsymbol{\alpha}_h) p(\boldsymbol{\alpha}_h)}{L(\mathbf{x}_i)} \left[\frac{1-\eta_{hj}}{f_{j0}(x_{ij})} \cdot \frac{\partial f_{j0}(x_{ij})}{\partial \phi_{j\eta}} + \frac{\eta_{hj}}{f_{j1}(x_{ij})} \cdot \frac{\partial f_{j1}(x_{ij})}{\partial \phi_{j\eta}} \right] \\ &= \sum_{i=1}^{I} \sum_{h=1}^{H} p(\boldsymbol{\alpha}_h|\mathbf{x}_i) \left[\frac{1-\eta_{hj}}{f_{j0}(x_{ij})} \cdot \frac{\partial f_{j0}(x_{ij})}{\partial \phi_{j\eta}} + \frac{\eta_{hj}}{f_{j1}(x_{ij})} \cdot \frac{\partial f_{j1}(x_{ij})}{\partial \phi_{j\eta}} \right]. \end{aligned} \tag{15}$$

By letting $p(\eta_j = \eta|\mathbf{x}_i) = \sum_{\{\alpha_h : \eta_{hj} = \eta\}} p(\boldsymbol{\alpha}_h|\mathbf{x}_i)$, where $\sum_{\eta=0}^{1} p(\eta_j = \eta|\mathbf{x}_i) = 1$, Equation 15 can be written as

$$\sum_{i=1}^{I} \left[p(\eta_j = 0|\mathbf{x}_i) \frac{1}{f_{j0}(x_{ij})} \frac{\partial f_{j0}(x_{ij})}{\partial \phi_{j\eta}} + p(\eta_j = 1|\mathbf{x}_i) \frac{1}{f_{j1}(x_{ij})} \frac{\partial f_{j1}(x_{ij})}{\partial \phi_{j\eta}} \right], \tag{16}$$

which reduces to

$$\sum_{i=1}^{I} p(\eta_j = \eta|\mathbf{x}_i) \frac{1}{f_{j\eta}(x_{ij})} \frac{\partial f_{j\eta}(x_{ij})}{\partial \phi_{j\eta}}, \tag{17}$$

because $\partial f_{j\eta^*}(x_{ij}) / \partial \phi_{j\eta} = \mathbf{0}$ when $\eta \neq \eta^*$.

Now, with Equation 3,

$$\frac{\partial f_{j\eta}(x_{ij})}{\partial \phi_{j\eta}} = \begin{pmatrix} \partial f_{j\eta}(x_{ij}) / \partial \mu_{j\eta} \\ \partial f_{j\eta}(x_{ij}) / \partial \sigma^2_{j\eta} \end{pmatrix} = \frac{f_{j\eta}(x_{ij})}{2\sigma^2_{j\eta}} \begin{pmatrix} 2(\log x_{ij} - \mu_{j\eta}) \\ (\log x_{ij} - \mu_{j\eta})^2 / \sigma^2_{j\eta} - 1 \end{pmatrix}. \tag{18}$$

Substituting Equation 18 into Equation 17 yields

$$\frac{\partial l(\mathbf{X})}{\partial \phi_{j\eta}} = \frac{1}{2\sigma^2_{j\eta}} \begin{pmatrix} 2\sum_{i=1}^{I} p(\eta_j = \eta|\mathbf{x}_i)(\log x_{ij} - \mu_{j\eta}) \\ \sum_{i=1}^{I} p(\eta_j = \eta|\mathbf{x}_i)[(\log x_{ij} - \mu_{j\eta})^2 / \sigma^2_{j\eta} - 1] \end{pmatrix}. \tag{19}$$

Finally, solving $\partial l(\mathbf{X})/\partial\phi_{j\eta} = \mathbf{0}$ for $\mu_{j\eta}$ and $\sigma^2_{j\eta}$ will yield the estimators

$$\hat{\mu}_{j\eta} = \sum_{i=1}^{I} p_{ij}(\eta)\log x_{ij}, \tag{20}$$

and

$$\hat{\sigma}^2_{j\eta} = \sum_{i=1}^{I} p_{ij}(\eta)\left(\log x_{ij} - \hat{\mu}_{j\eta}\right)^2, \tag{21}$$

where $p_{ij}(\eta) = p(\eta_j = \eta|\mathbf{x}_i)/\sum_{i=1}^{I} p(\eta_j = \eta|\mathbf{x}_i)$.

Computing the SEs. The *SE*s of the estimated item parameters can be calculated from the information matrix, $\boldsymbol{I}(\boldsymbol{\Phi}) = -E\{\partial^2 l(\mathbf{X})/\partial\boldsymbol{\Phi}^2\}$. The second derivative within this function can be carried out separately by the second-order partial derivative of the log-marginalized likelihood with respect to the parameters $\phi_{j\eta}$ and $\phi_{j^*\eta^*}$, which is given by

$$\frac{\partial^2 l(\mathbf{X})}{\partial\phi_{j\eta}\partial\phi_{j^*\eta^*}} = \sum_{i=1}^{I}\left\{\frac{1}{L(\mathbf{x}_i)}\frac{\partial^2 L(\mathbf{x}_i)}{\partial\phi_{j\eta}\partial\phi_{j^*\eta^*}} - \frac{1}{L^2(\mathbf{x}_i)}\left[\frac{\partial L(\mathbf{x}_i)}{\partial\phi_{j\eta}}\right]\left[\frac{\partial L(\mathbf{x}_i)}{\partial\phi_{j^*\eta^*}}\right]'\right\}. \tag{22}$$

Because the first term vanishes after the expectation is taken, Equation 22 becomes

$$\frac{\partial^2 l(\mathbf{X})}{\partial\phi_{j\eta}\partial\phi_{j^*\eta^*}} = -\sum_{i=1}^{I}\left[\frac{1}{L(\mathbf{x}_i)}\frac{\partial L(\mathbf{x}_i)}{\partial\phi_{j\eta}}\right]\left[\frac{1}{L(\mathbf{x}_i)}\frac{\partial L(\mathbf{x}_i)}{\partial\phi_{j^*\eta^*}}\right]'. \tag{23}$$

It has been shown in the last section that Equation 12 can be reduced to Equation 17, which is equivalent to say

$$\frac{1}{L(\mathbf{x}_i)}\frac{\partial L(\mathbf{x}_i)}{\partial\phi_{j\eta}} = p(\eta_j = \eta|\mathbf{x}_i)\frac{1}{f_{j\eta}(x_{ij})}\frac{\partial f_{j\eta}(x_{ij})}{\partial\phi_{j\eta}}. \tag{24}$$

Substituting Equation 18 into Equation 24 and thus to Equation 23 yields

$$\frac{\partial^2 l(\mathbf{X})}{\partial\phi_{j\eta}\partial\phi_{j^*\eta^*}} = -\sum_{i=1}^{I}\frac{p(\eta_j = \eta|\mathbf{x}_i)p(\eta_{j^*} = \eta^*|\mathbf{x}_i)}{4\sigma^2_{j\eta}\sigma^2_{j^*\eta^*}}\mathbf{d}_{j\eta}\mathbf{d}'_{j^*\eta^*}, \tag{25}$$

where

$$\mathbf{d}_{j\eta} = \begin{pmatrix} 2(\log x_{ij} - \mu_{j\eta}) \\ (\log x_{ij} - \mu_{j\eta})^2/\sigma^2_{j\eta} - 1 \end{pmatrix}. \tag{26}$$

Instead of computing the expectation, the information matrix is usually approximated using the observed $\mathbf{X}$, which results in $\boldsymbol{I}(\widehat{\boldsymbol{\Phi}})$. Note that the first term in Equation 22 will not be exactly equal to zero when using the observed data, particularly when the sample size is small. In this particular work, the effect

appears to be negligible, as demonstrated by the simulation study results. To calculate the *SE*s, Equation 25 is evaluated at $\hat{\phi}_{j\eta}$ and $\hat{\phi}_{j^*\eta^*}$, which yields the empirical information matrix. The inverse of the empirical information matrix provides an approximation of $\text{Cov}(\widehat{\mathbf{\Phi}})$, the root of which diagonal elements gives an approximation of the *SE*s of estimated item parameters.

References

Ben-Simon, A., Budescu, D. V., & Nevo, B. A. (1997). Comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, *21*, 65–88.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, *46*, 443–459.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley.

de Finetti, B. (1965). Method for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical & Statistical Psychology*, *18*, 87–123.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343–362.

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., & Chiu, C.-Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, *81*, 253–273.

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

Dressel, P. L., & Schmidt, J. (1953). Some modifications of the multiple choice item. *Educational and Psychological Measurement*, *13*, 574–595.

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, *37*, 655–670.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement*, *31*, 525–543.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston, MA: Kluwer Nijhoff.

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

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.

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.

Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, *77*, 615–633.

Meng, X.-B., Tao, J., & Chang, H.-H. (2015). A conditional joint modeling approach for locally dependent item responses and response times. *Journal of Educational Measurement*, *52*, 1–27.

Meng, X.-B., Tao, J., & Shi, N.-Z. (2014). An item response model for Likert-type data that incorporates response time in personality measurements. *Journal of Statistical Computation and Simulation*, *84*, 1–21.

Morin, C., & Bushnell, M. C. (1998). Temporal and qualitative properties of cold pain and heat pain: A psychophysical study. *Pain*, *74*, 67–73.

Noel, Y. (2014). A beta unfolding model for continuous bounded responses. *Psychometrika*, *79*, 647–674.

Noel, Y., & Dauvier, B. (2007). A beta item response model for continuous bounded responses. *Applied Psychological Measurement*, *31*, 47–73.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ranger, J., & Kuhn, J.-T. (2012). Improving item response theory model calibration by considering response times in psychological tests. *Applied Psychological Measurement*, *36*, 214–231.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, *34*, 1–97.

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*, 203–219.

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika*, *39*, 111–121.

Sie, H., Finkelman, M. D., Riley, B., & Smits, N. (2015). Utilizing response times in computerized classification testing. *Applied Psychological Measurement*, *39*, 389–405.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*, 481–520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*, 1–84.

Siegler, R. S. (1989). Hazards of mental chronometry: An example from children's subtraction. *Journal of Educational Psychology*, *81*, 497–506.

Tatsuoka, K. K. (1983). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

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

Thissen, D., Steinberg, L., Pyszczynski, T., & Greenberg, J. (1983). An item response theory for personality and attitude scales: Item analysis using restricted factor analysis. *Applied Psychological Measurement*, *7*, 211–226.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, *31*, 181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*, 287–308.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*, 5–20.

van der Linden, W. J. (2009). Predictive control of speededness in adaptive testing. *Applied Psychological Measurement*, *33*, 25–41.

van der Linden, W. J., Breithaupt, K., Chuah, S. C., & Zhang, Y. (2007). Detecting differential speededness in multistage testing. *Journal of Educational Measurement*, *44*, 117–130.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*, 365–384.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York, NY: Springer.

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for speededness in computerized adaptive testing. *Applied Psychological Measurement*, *23*, 195–210.

van der Linden, W. J., & van Krimpen-Stoop, E. M. L. A. (2003). Using response times to detect aberrant response patterns in computerized adaptive testing. *Psychometrika*, *68*, 251–265.

van der Linden, W. J., & Xiong, X. (2013). Speededness in adaptive testing. *Journal of Educational and Behavioral Statistics*, *38*, 418–438.

van der Maas, H. L. J., & Jansen, B. R. J. (2003). What response times tell of children's behavior on the balance scale task. *Journal of Experimental Child Psychology*, *85*, 141–177.

van der Maas, H. L., & Wagenmakers, E. J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, *118*, 29–60.

Wang, T., & Zeng, L. (1998). Item parameter estimation for a continuous response model using an EM algorithm. *Applied Psychological Measurement*, *22*, 333–344.

Authors

NATHAN D. MINCHEN is a PhD candidate at Rutgers, The State University of New Jersey, Room 304, 10 Seminary Place, New Brunswick, NJ 08901. His research interests include cognitive diagnosis modeling, item response theory, and computerized adaptive testing.

JIMMY DE LA TORRE is a professor at The University of Hong Kong, Room 520, Meng Wah Complex, Pokfulam Road, Hong Kong; email: j.delatorre@hku.hk. His research interests include item response theory, cognitive diagnosis modeling, and the use of diagnostic assessment to support classroom teaching and learning.

YING LIU is a research scientist at the University of Southern California, 635 Downey Way, Los Angeles, CA 90089, USA; email: liu385@usc.edu. Her research interests are psychometrics and behavioral statistics.

Manuscript received January 27, 2016
Revision received November 23, 2016
Accepted February 15, 2017