JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Winter 2009, Vol. 46, No. 4, pp. 450–469

Impact of Diagnosticity on the Adequacy of Models for Cognitive Diagnosis under a Linear Attribute Structure: A Simulation Study

Jimmy de la Torre
Rutgers, The State University of New Jersey
Tzur M. Karelitz
Education Development Center

Compared to unidimensional item response models (IRMs), cognitive diagnostic models (CDMs) based on latent classes represent examinees' knowledge and item requirements using discrete structures. This study systematically examines the viability of retrofitting CDMs to IRM-based data with a linear attribute structure. The study utilizes a procedure to make the IRM and CDM frameworks comparable and investigates how estimation accuracy is affected by test diagnosticity and the match between the true and fitted models. The study shows that comparable results can be obtained when highly diagnostic IRM data are retrofitted with CDM, and vice versa, retrofitting CDMs to IRM-based data in some conditions can result in considerable examinee misclassification, and model fit indices provide limited indication of the accuracy of item parameter estimation and attribute classification.

Most large-scale educational assessments are constructed based on the conventional unidimensional framework of item response theory (IRT), and are subsequently analyzed using unidimensional item response models (IRMs). Scores derived from this framework are useful in scaling and ordering students along a proficiency continuum, but these proficiency scores contain limited diagnostic information necessary for the identification of students' specific strengths and weaknesses. (Classical test theory is also used for scaling purposes, but this article focuses on the IRT framework because it subsumes the models that will be covered in our study.) Consequently, the single scores from these assessments may be of limited value in practical instructional settings. For assessments to help inform classroom instruction and learning, they must be cognitively diagnostic in that they must provide information that is "interpretative, diagnostic, highly informative, and potentially prescriptive" (Pellegrino, Baxter, & Glaser, 1999, p. 335).

Designing assessments to be more cognitively diagnostic requires a careful and thoughtful construction process. Equally important is the use of appropriate tools to extract relevant information from such assessments. One example of these tools is cognitive diagnosis models (CDMs). In contrast to traditional measurement models, CDMs are employed specifically for the purpose of diagnosing multiple finer-grained proficiencies (also referred to as attributes). In their simplest applications, instead of a single proficiency score, CDMs generate a binary profile to indicate which attributes are present or absent in each student. Because of their finer-grained nature, a profile provides specific information about the student's state of learning or understanding, and is relevant for subsequent actions, such as tailored instruction or remediation.

Despite the potential advantages of CDM-based score profiles, their benefits in practical educational settings have yet to be fully realized. One reason can be traced to the dearth of assessments constructed from a cognitive diagnosis framework (i.e., assessments specifically designed to measure multiple theory-based attributes). Consequently, many applications that employ CDMs rely on assessments developed based on a unidimensional IRM framework (i.e., assessments designed to measure a single continuous proficiency). This approach is sometimes referred to as *retrofitting* because it involves fitting CDMs to existing assessments *post hoc*. Recent examples of this approach include the analyses of the 2003 TIMSS data by Birenbaum, Tatsuoka, and Xin (2005), and Tatsuoka, Corter, and Tatsuoka (2004), the 2003 NAEP data by de la Torre (2006), and the GRE data by Gorin and Embretson (2006). Other retrofitting applications of large-scale assessments are described in Roussos, Templin, and Henson (2007).

Retrofitting assumes that CDMs can be employed to extract diagnostic information from an assessment irrespective of the framework used in its construction. Thus far, the validity of this assumption has only been examined informally, primarily through model-data fit analyses, but not directly. To support such a practice, it is imperative that the validity of the assumption underlying retrofitting be examined empirically. In this study, we will use simulated data to systematically examine the extent to which data from unidimensional IRM-based assessments can be analyzed using CDMs. In addition, this study will also examine the complementary assumption that data from CDM-based assessments can be analyzed using unidimensional IRMs.

Background

IRM versus CDM. Although cognitive diagnosis can be performed using other approaches (e.g., analysis of error patterns such as the rule space theory; Tatsuoka, 1983), this article focuses on approaches that are based on psychometric models that analyze data at the individual item response level, and can be employed for diagnostic purposes (i.e., IRT-based CDMs; DiBello & Stout, 2007). Thus, we differentiated between two classes of psychometric models: one class consists of models that are specifically developed to provide finer-grained information; another class consists of models that are not originally intended, but nonetheless can be used for diagnostic ends. In this article, the former will be represented by the *deterministic inputs, noisy “and”* gate (DINA; Junker & Sijtsma, 2001) model, and the latter by the two-parameter logistic (2PL) model. The two classes of IRT-based CDMs can be further distinguished from one another based on the number and nature of the underlying latent variables these models assume: the former assumes *multiple discrete* latent attributes, while the latter assumes *a single continuous* latent proficiency. Additional discussion on the distinction of IRT-based CDMs based on the underlying latent variable can be found in Roussos et al. (2007), and Stout (2007).

It should be noted that although unidimensional IRMs such as the 2PL model can be used to extract diagnostic information under certain conditions, they are typically and primarily used for a different (i.e., scaling) purpose. In contrast, the DINA model is developed and used primarily, if not exclusively, for diagnostic purposes. Based on the difference in their primary usage, we chose to classify the 2PL as a (unidimensional) IRM, and the DINA model as an (IRT-based) CDM. Henceforth, when no additional qualifications are given, the 2PL model will be used interchangeably with IRM, and the DINA model with CDM.

Diagnosis with unidimensional IRMs. As stated earlier, unidimensional IRMs, although based on a single latent trait, can be used to provide more diagnostic information. In such applications, scale scores are accompanied with exemplars or descriptions of the type of problems students at different levels of proficiency can do. For example, in the 2007 National Assessment of Education Progress (NAEP) 4th grade mathematics assessment, students who score about 214 are located on the proficiency continuum close to items that require students to *identify congruent triangles*, and *determine the fraction of a figure that is shaded* (Lee, Grigg, & Dion, 2007). A student's location relative to the items can indicate the types of problems the student has and has not mastered.

Another example of how IRM model can be employed for diagnostic purposes is the Rasch measurement model framework (Rasch, 1960), which has been used as a basis for configuring the proficiency into a linear (hierarchical) structure. Under this framework, a unidimensional continuous trait represents a theory-based ability structure composed from a set of knowledge, skills or behaviors ordered by difficulty levels. Thus, a trait's linear structure is simply a reflection of the linear structure of the knowledge, skills or behavior defining the trait. This structure is also referred to as a *construct map*—an underlying continuum on which we can order things (items, persons, responses) with respect to a substantively defined construct (Wilson, 2005). Thus, a construct map allows a continuous trait to be mapped onto discrete categories that have a linear structure.

DINA model. In investigating how data generated within an IRM framework can be analyzed using a CDM framework, and vice versa, we assume the readers are familiar with the common unidimensional IRMs, specifically the 2PL model. The following section reviews how CDMs represent items and examinees using discrete attributes in the context of the DINA model. The DINA model is a discrete latent variable CDM that has been the foundation of several approaches to cognitive diagnosis and assessment (e.g., Doignon & Falmagne, 1999; Haertel, 1990; Tatsuoka, 1995). It is a conjunctive model in that it assumes that all the required attributes are necessary for successful completion of the item, and it is also a noncompensatory model in that the absence of one attribute cannot be compensated by the presence of other attributes (see de la Torre, 2009; Maris, 1999; Roussos et al., 2007). The model combines a deterministic component, which represents how students are *expected* to perform on each item, with a probabilistic component, which represents how they are *likely* to perform. As such, the DINA is a simple but interpretable model. Despite its simplicity, the DINA model has been shown to provide good model-data fit (e.g., de la Torre & Douglas, 2008).

Like many CDMs, implementation of the DINA model requires the construction of a $J \times K$ binary matrix called the Q-matrix (Tatsuoka, 1983) to specify how each item is related to each of the attributes. The J and K represent the numbers of items and attributes, respectively. An element of the matrix q_{jk} is equal to 1 if the *k*th attribute is required to correctly answer the *j*th item; otherwise, it is 0. Successful completion of an item may require one or more attributes. The Q-matrix plays an important role in test development in that it embodies the attribute blueprint or cognitive specifications for test construction (Leighton, Gierl, & Hunka, 2004; Junker, 1999).

In the formulation of the DINA model, we denote the vector of dichotomous item responses of student i to J items by $\boldsymbol{Y}_i$. The vector of responses is a function of K attributes specified for the test. Let $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}'$ represent the student *i*'s attribute vector. The *k*th element of the vector, α_{ik}, is 1 when student i possesses the *k*th attribute; otherwise, it is equal to 0. Given the *i*th student's attribute vector, $\boldsymbol{\alpha}_i$, and the *j*th row of the Q-matrix, the DINA model generates the latent ideal response, η_{ij}, through the deterministic function $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. Hence, the model classifies respondents into two latent groups for each item: students who possess all attributes required for the item (and thus expected to respond correctly), and students who lack at least one of the required attributes (and thus expected to respond incorrectly). The noisy component of the model allows the possibility for students who possess all the required attributes for item j to slip and answer the item incorrectly, and students who do not possess all the required attributes to guess and answer item j correctly. Finally, given the slip and guessing parameters s_j and g_j, the item response function is written as

$$P(Y_{ij} = 1 \,|\, \boldsymbol{\alpha}_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \tag{1}$$

For additional discussions and applications of the DINA model, see Doignon and Falmagne (1999), Haertel (1990), Junker and Sijtsma (2001), Macready and Dayton (1977), Maris (1999), and Tatsuoka (1995). Estimation of the model can be found in de la Torre (2009).

Latent class structures. Each attribute pattern represents a specific combination of attribute mastery and nonmastery, and therefore can be viewed as a unique latent class. In a domain with K attributes there are 2^K distinct latent classes. Although many applications of cognitive diagnosis are based on the assumption that attributes are independent, cognitive research suggests that cognitive skills should not be investigated in isolation (Kuhn, 2001; Tatsuoka, 1995). In most applications, it is more reasonable to assume that attributes, and thus latent classes, are dependent and follow some type of structure.

In general, latent classes are considered structured if some constraints exist regarding the relationships among the attributes. If the attributes are related such that the mastery of an attribute is a prerequisite to the mastery of another, the structure of the latent classes is said to be hierarchical. Leighton et al. (2004) presented different types of hierarchical attribute structures. For example, within the linear structure, mastery of simpler attributes is a prerequisite to the mastery of more complex attributes. The hierarchical structure simplifies the space of the latent classes. With five attributes, instead of $2^5 = 32$ classes, only six are possible under this structure, namely $\{00000\}$, $\{10000\}$, $\{11000\}$, $\{11100\}$, $\{11110\}$, and $\{11111\}$.

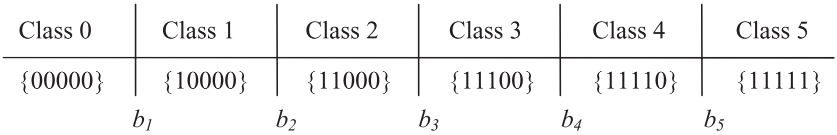

FIGURE 1. *A linear hierarchical structure for 5 attributes and 6 classes.*

Present Study

The goal of this study is to investigate the ability of measurement models to retrieve diagnostic information as a function of the method by which the data were generated. Specifically, we considered two factors that are important in practical situations: the diagnostic power of the test items, and the match between the true and fitted models. We expected that more diagnostic items and a better match between true and fitted models would improve estimation accuracy of model parameters.

This study was designed to directly evaluate the validity of the retrofitting assumption under the setup of a hierarchical linear attribute structure. We chose this particular structure for the following reasons. First, a set of discrete attributes with a linear structure is used because it provides a good correspondence to the unidimensional latent trait. In Figure 1, the five cut points b_1 to b_5 represent locations on the θ scale at which students acquire mastery of each attribute (students in Class 0 have not mastered any attributes). This dual representation makes results from continuous (IRM) and discrete (CDM) representations comparable. In contrast, it can be shown that other attribute structures (e.g., divergent or convergent structures) would require multiple latent traits for a similar correspondence to be established. That is, more complex cognitive structure may require a multidimensional IRM, which is beyond the scope of this article.

In addition to its implicit use in many applied settings, the linear attribute structure can also be used in domains where acquisition of knowledge, skills or competencies can be viewed as developmental (e.g., Commons & Richards, 1984; Erikson, 1950; Inhelder & Piaget, 1958), or where mastery of objectives can be represented by a hierarchical taxonomy (e.g., Anderson & Krathwohl, 2001).

Consequently, our main research question focuses on how well discrete models (i.e., CDMs) fit data from a continuous process (i.e., data obtained under an IRM framework) as a function of the diagnosticity of the data (see definition below) under the linear attribute structure. Additionally, we are interested in how estimation accuracy varies across subgroups of parameters. That is, how accuracy might vary from one ability class or item class to another as the conditions change. Our secondary research question is the complement of CDM retrofitting—how well do IRMs fit data from a discrete process? This question may have less practical significance because tests designed in a CDM framework are usually analyzed using CDMs, yet it can provide some important insights concerning the extent to which IRMs can extract diagnostic information that exists in the data.

In this study, we used simulated data to control the sources of systematic and random variation observed in assessment data. The process of extracting diagnostic information from assessments assumes that the data contain both diagnostic information and random error. In simulation studies, the *diagnosticity* of the generated data depends on how these information and error components are represented in the generating model. We define diagnosticity as the extent to which the data can be adequately represented by underlying discrete structures. Assessment information is most useful for diagnostic purposes when the process that generated the data can be well represented by discrete categories. In other words, when the diagnostic component is larger relative to the error component, data with greater diagnosticity can be expected. The diagnostic component is larger than the error component when the expected response has small variability (i.e., the probability of success is close to either zero or one) across the examinees. We would like to note that our definition of diagnosticity does not preclude the possibility of generating data with high diagnosticity using IRMs, nor does it assume that CDMs necessarily result in diagnostic data. For example, highly discriminating items with difficulty parameters located at a few specific points can produce data with a desirable level of diagnosticity. For data obtained under a CDM framework, diagnosticity may be low if the items' error probabilities (i.e., slip and guessing) are high.

Simulation Study

Design

This simulation study was designed to manipulate the item characteristics that contribute to test diagnosticity. Different measurement models (i.e., IRM and CDM) were used to both generate and analyze the data. Our goal was to investigate how the congruence or lack thereof between the generating and fitted model affects the estimation accuracy under the different conditions of diagnosticity. Specific actions were taken to increase the comparability of data sets across the study conditions.

The study focused on a generic test design of multiple items targeting the same latent ability with varying item difficulties and discriminations. The Q-matrix in Table 1 represents the structure of the test that was designed based on a linear structure of 5 attributes (as explained in Figure 1). The linear structure defines 5 item classes, and within each item class 5 items were generated to create a 25-item test. Each item class can be unidimensionally represented as a cut point on the latent continuum. Here, the cut points were chosen to be −2, −1, 0, 1, and 2. The item

TABLE 1
Q-Matrix Representing the Hierarchical Structure in the Simulation Study

		Attributes				
Item Class	Item Numbers	1	2	3	4	5
1	1–5	1	0	0	0	0
2	6–10	1	1	0	0	0
3	11–15	1	1	1	0	0
4	16–20	1	1	1	1	0
5	21–25	1	1	1	1	1

difficulties within the item class c were selected to be equally spread within ±.05 of the cut point b_c. In this setup, the K attributes implied $K+1$ ability classes, but only K item classes.

We generated the data by manipulating three factors. The first factor was the diagnosticity of the test. Because this study considered a unidimensional trait and a linear attribute structure, diagnosticity can be represented by the item discrimination of the measurement model. We acknowledge that there are other factors that can contribute to the diagnosticity of a particular item. However, we chose the discrimination parameter to control the diagnosticity of the simulated data because of its usefulness in separating the performance of examinees with abilities below and above the cut points. We initially created item parameters for a 2PL model and then converted them to the corresponding parameters in a DINA model, as explained below. Three diagnosticity conditions were generated. The extreme conditions were created using discrimination parameters of $a \sim U(.4, .8)$ for the low condition and $a \sim U(1.6, 2.0)$ for the high condition. The medium diagnosticity condition was designed to mimic a typical test. For this condition, the discrimination parameters were taken from a pool of 550 nationally standardized items. The parameters, which ranged from .4 to 2.0, were stratified to ensure similar diagnosticity across the item classes.

The second factor manipulated in this study was the type of generating or "true" model. In one condition we generated data assuming the true model is IRM-based (i.e., 2PL model), and in the other condition we assumed the true model is CDM-based (i.e., DINA model). To make the two data sets comparable, we began by generating the IRM parameters and converted them into the CDM parameters. The details of this conversion are described in the next section.

The third factor was the type of estimating or fitted model. All data sets were analyzed using both the 2PL model and the DINA model. Therefore, the study design included 12 conditions; four conditions for each diagnosticity level. There were two retrofitting conditions: retrofitting the DINA model to IRM data, and the 2PL to CDM data. There were two baseline conditions: fitting the DINA model to CDM data, and the 2PL to IRM data. Within each true model-diagnosticity condition, 100 data sets with $I = 5{,}000$ examinees were generated and estimated using the two models. To fit the 2PL model we used BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996), and to fit the DINA model we used an EM algorithm written in Ox (Doornik, 2003). The 2PL analysis was done without assuming *a priori* structure on examinee distribution (i.e., using an empirical prior); the DINA model was implemented without constraints on the attribute structure.

Parameter conversion from IRM to CDM. Parameter conversion from IRM to CDM served two purposes. First, it allowed comparable data to be generated using different representational frameworks, and second, it provided a baseline with which results from retrofitting the DINA model to IRM data can be compared. Initially, we generated the examinee and item parameters for the IRM, and then converted them to the CDM parameters. The ability parameters for the IRM were sampled from $N(0, 1)$. The abilities were deterministically transformed into the corresponding ability classes based solely on the location of the cut points. For example, abilities below the first cut point (-2) were placed in ability class 0, abilities between the first and second cut points (-2 and -1) were placed in class 1, and so on. The final outcome of this conversion procedure was 100 sets of 5,000 examinees' θs (used to generate data for the 2PL) and their corresponding deterministic ability classes (used to generate data for the DINA model).

To transform the 2PL item parameters into the DINA model's slip and guessing parameters we introduced a procedure called the logistic-to-step transformation (LST). This procedure uses the item class cut point to divide the examinee distribution into two groups: masters (above the cut point) and nonmaster (below the cut point). A group's probability of correct response to the item is based on the region of the item characteristic curve (ICC) and the ability density specific to the group. LST transforms the 2PL model parameters into the DINA model parameters using the group-level expected misclassification indexes (G-EMIs). With respect to item j in class c, the expected misclassification indices for examinees in groups $\eta_j = 0$ and $\eta_j = 1$ are defined as

$$G\text{-}EMI_j^{(0)} = \int_{-\infty}^{b_c} P_j(\theta)g(\theta)d\theta \bigg/ P_j^{(*)}(b_c),$$

and

$$\begin{aligned} G\text{-}EMI_j^{(1)} &= \int_{b_c}^{\infty} \left[1 - P_j(\theta)\right] g(\theta)d\theta \big/ \left[1 - P_j^{(*)}(b_c)\right] \\ &= 1 - \int_{b_c}^{\infty} P_j(\theta)g(\theta)d\theta \big/ \left[1 - P_j^{(*)}(b_c)\right], \end{aligned}$$

respectively, where $P_j(\theta)$ is the 2PL function, b_c is the cut point, and $P_j^{(*)}(b_c) = \int_{-\infty}^{b_c} g(\theta)d\theta$ is the proportion of examinees below the cut point. In other words, $G\text{-}EMI_j^{(0)}$ represents the weighted probability of examinees in group $\eta_j = 0$ answering item j correctly; in contrast, $G\text{-}EMI_j^{(1)}$ represents the weighted probability of examinees in group $\eta_j = 1$ responding incorrectly to item j. Thus, using LST, the resulting DINA model parameters are $g_j = G\text{-}EMI_j^{(0)}$ and $s_j = G\text{-}EMI_j^{(1)}$, respectively. We used the standard normal distribution as the $g(\theta)$ to compute the DINA model parameters. LST can also be applied to other unidimensional IRMs (e.g., 1PL, 3PL), and can be carried out using indices other than G-EMIs.

Table 2 lists the 2PL and DINA item parameters under the three diagnosticity conditions, where items are grouped by class (i.e., 2PL item difficulty). Within each item class we present the minimum and maximum values for the different item parameters. Note that as the diagnosticity of the data increased, the discrimination, slip and guessing parameters changed, such that the probability of correct response for most examinees approached 0 or 1.

Parameter estimation accuracy. To study the accuracy of model parameter estimation, different statistics were employed depending on the variable type. For continuous variables (e.g., s and g), the root mean squared error (RMSE) was calculated across the 100 replications. The effect of diagnosticity on estimation accuracy was

TABLE 2
Item Parameters for Simulation Study

Item			Discrimination			Slip			Guessing		
Class	Range	Difficulty	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
1	Min	−2.05	.42	.42	1.66	.10	.02	.02	.39	.30	.25
	Max	−1.95	.78	1.81	1.92	.20	.21	.03	.44	.43	.31
2	Min	−1.05	.45	.44	1.81	.21	.09	.07	.36	.22	.20
	Max	−.95	.71	1.77	1.98	.29	.28	.09	.41	.41	.25
3	Min	−.05	.41	.42	1.63	.29	.16	.14	.29	.16	.14
	Max	.05	.76	1.76	1.95	.36	.36	.19	.38	.37	.17
4	Min	.95	.45	.43	1.75	.33	.21	.19	.20	.09	.07
	Max	1.05	.79	1.83	1.97	.40	.41	.25	.29	.29	.09
5	Min	1.95	.42	.42	1.63	.39	.30	.26	.12	.02	.02
	Max	2.05	.69	1.82	1.96	.44	.43	.31	.21	.21	.03

Notes Med. = medium, Diag. = diagnosticity, Min = minimum, Max = maximum.

calculated across all the parameters as well as within subgroups of the parameters. For discrete variables (i.e., attribute vector $\boldsymbol{\alpha}$) we calculated the proportions of misclassification for each attribute and for the whole attribute vector. Let α_{ik} be the true classification of attribute k for examinee i and $\hat{\alpha}_{ik}$ the estimated attribute classification. The proportion of misclassification for attribute k and vector misclassifications are computed as

$$\frac{\sum_i |\alpha_{ik} - \hat{\alpha}_{ik}|}{I}$$

and

$$\frac{\sum_i \left[1 - \prod_k (1 - |\alpha_{ik} - \hat{\alpha}_{ik}|)\right]}{I},$$

respectively. As in the continuous variables, the percent of misclassification can be calculated for the whole parameter set or within subgroups of the parameters (i.e., for each attribute or the whole attribute vector within an ability class).

Finally, when CDM-generated data were analyzed with the 2PL model, the estimated parameters were compared to the original 2PL parameters (i.e., parameters *before* the conversion procedures were applied to obtain the DINA parameters). In addition, the estimated 2PL parameters were converted using the same procedures and the converted parameters were compared to the CDM generating parameters. In contrast, when IRM-generated data were analyzed with the DINA model, the estimated parameters were only compared to the DINA parameters, which were obtained *after* the conversion procedures.

Results

Overall model fit. Table 3 gives the fit statistics when the 2PL and DINA models were fitted to the data generated under the different conditions. Both the root mean squared deviation (RMSD) and the –2× log-likelihood led to the same conclusions. Here RMSD is defined as $\sqrt{\sum_{i=1}^{I} \sum_{j=1}^{J} (X_{ij} - E_{ij})^2}$, where X_{ij} and E_{ij} are the observed and model-based responses, respectively. In general, better model-data fit (fit statistics with lower values) was obtained when the items had high diagnosticity regardless of the fitted model. This suggests that under very specific conditions (i.e., high diagnosticity, linear hierarchy structure), retrofitting the DINA model to the IRM data may be justified given that the model-data fit statistics remained relatively small. However, additional improvement in the model fit can be obtained when correct models are used (i.e., the fitted models matched the true models), regardless of the nature of the underlying latent trait (continuous or discrete). Finally, in addition to the correct models having better model fit, the differences between the model fit statistics of the correct and retrofitted models were smaller compared to other conditions under the low diagnosticity condition. This indicates that the consequences of

TABLE 3
Fit Statistics across True and Fitted Models

		True Model					
		IRM			CDM		
Fit Statistic	Fitted Model	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
RMSD	2PL	138.7	116.4	83.4	147.9	129.9	103.6
	DINA	140.5	120.5	90.5	144.4	125.0	96.1
−2 × Log- Likelihood	2PL	117,342	84,760	46,086	132,113	106,491	74,854
	DINA	119,580	89,879	54,611	128,030	100,790	66,912

Note. Med. = medium, Diag. = diagnosticity.

TABLE 4
Item Parameter RMSE from DINA Analysis

		Guessing			Slip		
True Model	Item Class	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
IRM	1	.27	.35	.31	.10	.11	.06
	2	.13	.16	.07	.09	.16	.10
	3	.04	.09	.07	.04	.10	.07
	4	.09	.10	.11	.14	.15	.06
	5	.08	.10	.05	.27	.34	.33
CDM	1	.08	.10	.07	.05	.10	.03
	2	.04	.11	.07	.05	.13	.06
	3	.05	.09	.06	.05	.09	.05
	4	.05	.06	.06	.04	.06	.08
	5	.05	.09	.03	.08	.09	.07
Overall IRM		.15	.19	.16	.15	.19	.16
Overall CDM		.05	.09	.06	.05	.10	.06

Note. Med. = medium, Diag. = diagnosticity.

using an incorrect model were smaller when the items were less diagnostic. In other words, the noise introduced by low diagnosticity items was more influential on estimation inaccuracy than the noise introduced by fitting the wrong model.

Item and person estimation accuracy for the DINA analysis. Given in Table 4 are the RMSEs of the DINA parameter estimates for both the IRM and CDM data by item class. The results in Table 4 show three interesting patterns. First, as expected, fitting DINA to CDM data provided more accurate results than retrofitting DINA to IRM data for both slip and guessing parameters. Second, within each data type, the pattern of estimation accuracy across item classes was reversed for the guessing and

TABLE 5
Percent of DINA Misclassification by Attribute and Vector

	True Model					
	IRM			CDM		
Misclassification	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
Attribute 1	30.3	18.5	8.4	1.6	.8	.4
Attribute 2	28.4	19.1	9.7	10.1	4.4	1.8
Attribute 3	15.2	11.0	8.3	19.7	8.0	2.9
Attribute 4	27.4	18.8	9.9	9.2	4.2	1.8
Attribute 5	30.4	18.0	8.8	1.6	.7	.4
Ability vector	84.8	66.8	41.8	35.4	16.6	7.1
One attribute	43.1	48.9	38.4	29.5	15.1	6.9

Note. Med. = medium, Diag. = diagnosticity.

slip parameters. This is most evident in the IRM data, where the worst accuracy for guessing was in the lowest item class and the worst accuracy for slip was in the highest item class. Inspection of the signed bias values revealed that in both cases the parameters were overestimated. In other words, retrofitting DINA to IRM data produced estimates that were too easy (difficult) for the low (high) item classes. Third, estimation accuracy did not monotonically increase with diagnosticity. In fact, the medium diagnosticity condition (i.e., a typical test) produced the highest parameter estimation inaccuracy for both the CDM and the IRM data, whereas the low and high diagnosticity conditions were relatively similar. The medium condition had a larger range of discrimination parameters compared to the two other conditions, which suggests that the variability of discriminations may have a stronger effect on parameter estimation accuracy compared to their actual discrimination power.

Table 5 gives the percent of misclassification by attribute based on the DINA analysis. The misclassification rate was computed by comparing the estimated classification against the deterministic classification obtained using the true abilities. The table shows that when the items were more diagnostic the misclassification rates were consistently lower for both the CDM and IRM data. In addition, except for attribute 3 under the low diagnosticity condition, the percent of misclassification was lower for the CDM data. For IRM, the extreme ability classes were badly classified compared to the middle classes, whereas for CDM data, the reverse pattern was observed. Overall, the attribute classification accuracy was considerably lower when the DINA model was applied to IRM data with low item diagnosticity. The second to last row of Table 5 shows the percent of vectors with at least one misclassified attribute. As a whole, the vector misclassification rate of the CDM data was much better than the IRM data across all diagnosticity levels. When diagnosticity was low the DINA analysis largely failed to recover the attribute vector for IRM-generated data, but was moderately successful with the CDM data.

The last row of Table 5 shows the percent of examinees with a single misclassified attribute. For examinees whose attribute vector was incorrectly estimated, the

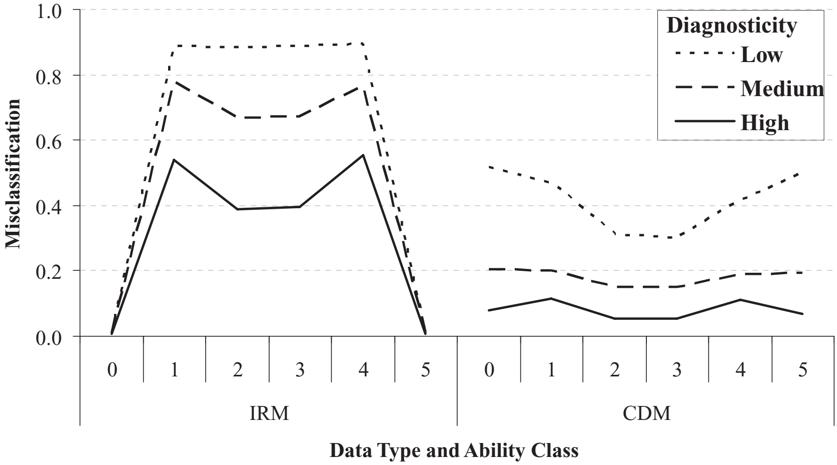

FIGURE 2. *Proportions of DINA vector misclassification by ability class.*

majority had only one attribute misclassified. The difference between the second to last and the last rows of Table 5 gives the percent of examinees with more than one misclassified attribute. As diagnosticity increased, the percent of more than one attribute misclassification dramatically decreased. In addition, an inspection of the misclassification patterns showed that attributes were symmetrically misclassified as mastered or nonmastered. For example, from the 43% of examinees who had one misclassified attribute in the IRM low diagnosticity case, 21% were attribute nonmasters who were classified as masters, whereas 22% were masters who were classified as nonmasters. The same pattern held across all diagnosticity levels and for both data types (these results are not shown in Table 5).

A closer inspection of the DINA analysis of the IRM and CDM data is provided by the vector misclassification rates of the six ability classes in Figure 2. The vector misclassification rate in Table 5 is the weighted average of the values shown in Figure 2. Overall, as diagnosticity increased, classification improved for all ability classes. Although the CDM data provided uniformly smaller misclassification rates than the IRM data across the classes, the pattern indicates that the difference between the models decreased as diagnosticity increased.

When diagnosticity was low, CDM data had more misclassifications in the extreme classes than the middle classes. However, as diagnosticity increased this difference disappeared; all ability classes were classified similarly well. The figure shows that when diagnosticity was low, retrofitting DINA to IRM data led to very high misclassification of most examinees in ability classes 1 through 4. For ability classes 0 and 5, the accuracy from the IRM data was better than the CDM data. Regardless of diagnosticity, these classes were very well classified. Yet, that hardly affects the overall accuracy because there were very few examinees in those classes.

Taken together, the DINA results suggest that the overall model-fit statistics can provide an indication of how accurately the examinee parameters can be estimated. Under the high diagnosticity condition where model fit was high, the examinee

TABLE 6
Item Parameter RMSE from 2PL Analysis

		Difficulty			Discrimination		
True Model	Item Class	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
IRM	1	.27	.15	.09	.07	.10	.13
	2	.14	.07	.04	.07	.08	.10
	3	.04	.03	.02	.06	.08	.09
	4	.14	.07	.04	.07	.08	.09
	5	.27	.14	.09	.07	.10	.13
CDM	1	8.13	5.07	2.43	.53	.92	1.34
	2	1.94	1.04	.52	.39	.75	1.12
	3	.08	.04	.03	.30	.56	.84
	4	1.63	1.00	.52	.42	.77	1.11
	5	7.84	5.20	2.50	.49	.93	1.31
Overall IRM		.19	.10	.06	.07	.09	.11
Overall CDM		5.20	3.33	1.60	.44	.80	1.16

Note. Med. = medium, Diag. = diagnosticity.

classifications were also very accurate, and under the low diagnosticity condition where model fit was low, parameters were less accurately estimated. Item parameter estimation seemed to be more strongly effected by the variability in diagnosticity, rather than its magnitude. This result was not reflected in the model-fit indices. Moreover, model-fit indices showed only a small advantage for having a match between the true and fitted model, whereas the results for item and examinee parameters showed that within the conditions of this study there were great discrepancies when the DINA model was retrofitted to IRM-generated data. The discrepancies were much smaller when the correct model was used. However, the effect of model-matching seemed to decrease as diagnosticity increased.

Item and person estimation accuracy for the 2PL analysis. Table 6 shows the RMSEs for the difficulty and discrimination parameters estimates from the 2PL analysis. Several interesting patterns can be observed. First, estimation accuracy was much better when there was a match between the true and fitted model. Estimation at some item classes could be very inaccurate when the 2PL model was retrofitted to CDM data. Second, accuracy improved with diagnosticity for the difficulty parameter, but deteriorated for the discrimination parameter. That is, highly diagnostic items can be accurately located on the continuous scale, but their slopes may not. An inspection of the signed bias revealed that discrimination was always underestimated. Third, estimation accuracy varied greatly across ability classes. Generally, parameters of items in the middle class were most accurately estimated, and items in the extreme classes were least accurately estimated, and this can be attributed to the number of examinees present in those ability classes. (In a pilot study, estimation accuracy did not change across the item classes when the ability distribution was uniform.) Inspection of the signed biased revealed that while the discrimination parameter was always

TABLE 7
Item Parameter RMSE from 2PL Analysis Converted to DINA Parameters

		Guessing			Slip		
True Model	Item Class	Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
IRM	1	.07	.07	.06	.01	.01	.00
	2	.04	.03	.03	.01	.01	.01
	3	.02	.01	.01	.02	.01	.01
	4	.01	.01	.01	.04	.03	.03
	5	.01	.01	.00	.07	.06	.06
CDM	1	.39	.47	.54	.01	.01	.01
	2	.21	.26	.27	.04	.04	.04
	3	.08	.09	.09	.08	.09	.09
	4	.04	.04	.04	.22	.26	.27
	5	.01	.01	.01	.37	.47	.54
Overall IRM		.04	.03	.03	.04	.03	.03
Overall CDM		.20	.24	.27	.20	.24	.27

Note. Med. = medium, Diag. = diagnosticity.

underestimated, the difficulty parameter was underestimated for low item classes and overestimated for high item classes.

The large inaccuracies in retrofitting CDM data with an IRM can be explained by the difference in the item response functions of the 2PL and DINA models: the former has 0 and 1 as its asymptote, whereas the latter has g and $1 - s$. To approximate the upper and lower asymptotes of the step function that generated the data, the logistic function had to be flatter (hence discrimination was underestimated) and more spread out (hence low item classes were underestimated and high item classes were overestimated).

Finally, the 2PL item estimates were converted to DINA parameters and compared with the CDM item parameters. These results are given in Table 7, which shows the RMSEs for the guessing and slip parameters. Table 7 can be compared with Table 4 since both correspond to the same parameters. Inspection of the two tables shows similar patterns and indicates that both analyses performed comparably with respect to these item parameters. In other words, the 2PL analysis of IRM data produced similar (small) biases compared to the DINA analysis of the CDM data, and the 2PL analysis of CDM data produced similar (relatively large) biases compared to the DINA analysis of the IRM data. A comparison of the Overall IRM row in Table 7 and Overall CDM in Table 4 shows that when the true and fitted models match, one could expect good recovery of item parameters, slightly better for the continuous case.

In this study, we also examined the extent to which IRMs can be useful for diagnostic purposes. In the current setup, the examinees' 2PL-estimated θs were converted into ability classes based on the cut points and the deterministic classification procedure. This classification was compared against the deterministic classification based on the true θ. Table 8 shows that for the 2PL analysis, ability estimation

TABLE 8
Examinee Parameters Estimation Accuracy for the 2PL Analysis

		True Model					
		IRM			CDM		
		Low Diag.	Med. Diag.	High Diag.	Low Diag.	Med. Diag.	High Diag.
	θ RMSE	.47	.33	.25	.82	.59	.47
Percent	Ability vector	36.6	26.0	19.0	52.1	32.4	17.3
Misclassification	One attribute	36.0	26.0	19.0	46.0	31.5	17.1

Note. Med. = medium, Diag. = diagnosticity.

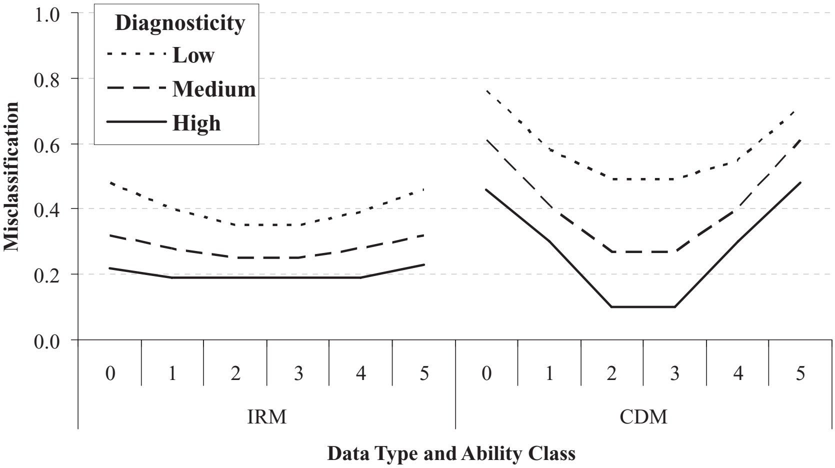

FIGURE 3. *Proportion of 2PL vector misclassification by ability class.*

improves with diagnosticity. Both RMSE (calculated based on the difference between the true and estimated θ) and the percent of ability class misclassification decreased as diagnosticity increased. Overall, the results show an advantage to having a match between the true and the fitted model, although when diagnosticity was high, the level of ability misclassification was about the same for the IRM and the CDM data. The last row of Table 8 shows the percent of ability misclassifications that involved only one misclassified attribute. Misclassifications involving more than one attribute were very rare. Inspection of the misclassification pattern indicates that the 2PL analysis classified examinees one ability class below or above their true class to the same degree.

The proportions of 2PL-derived attribute vector misclassifications by ability class are shown in Figure 3. The results indicate that within a particular data type (i.e., IRM or CDM), highly diagnostic items produced lower misclassifications overall. Estimation of IRM data produced similar misclassification rates across the 6 ability classes whereas CDM data produced worse classification rates for classes at the extreme ends. Under the low diagnosticity condition, both the IRM and CDM data were estimated with high levels of misclassifications, although the IRM results were uniformly better than the CDM results. In contrast, under the high diagnosticity condition, the CDM data had smaller misclassification rates than the IRM data for the middle ability classes (2 and 3).

The low diagnosticity results presented in Figures 2 and 3 show that the use of the correct model was crucial under this condition in minimizing the overall error of misclassification. Additionally, for attribute classification purposes, analyzing CDM data with an IRM produced less discrepancy from the true parameters than analyzing IRM data with a CDM. In other words, under the framework of this study, retrofitting discrete models to data generated by a continuous model was not supported unless the items are of high diagnosticity. Finally, comparing ability classification under the two model-matching situations (i.e., fitting DINA to CDM in the right panel of Figure 2 and fitting 2PL to IRM in the left panel of Figure 3) indicates that results were comparable under the low diagnosticity condition, but when diagnosticity was high, better classification was achieved when both the true and fitted models were discrete.

Discussion

The primary aim of this study was to examine how the congruence between the underlying model and fitted model affects DINA item parameter estimation and attribute classification under different diagnosticity conditions. Through a carefully designed simulation study where important factors were manipulated, and continuous and discrete representations were made comparable, we examined the viability of retrofitting CDMs to IRM test data when a unidimensional trait (represented as a hierarchical linear attribute structure) can be assumed to underlie the response process. The results of the simulation study indicate that CDM-analysis of IRM data may not be tenable particularly when the items are of low diagnosticity. That is, for the IRM data, the ability misclassification rate was remarkably poor under the low diagnosticity condition investigated in this study, and remained relatively unsatisfactory even when highly diagnostic items were involved. If anything, the results of the simulation study suggest that retrofitting CDM data with IRM is appropriate when the items are of typical or high diagnosticity. However, such an approach would still be considered suboptimal in comparison to the much smaller misclassification rates obtained using the correct CDM. These results underscore the importance of choosing a model appropriate to the nature of the data to obtain the best classification results. It should also be noted that these findings were obtained from highly comparable IRM and CDM representations. Thus, we can expect a higher degree of misclassification when more complex structures are involved and the correspondence between the two representations are far from ideal.

Although the model-fit statistics included in this study serve several uses (e.g., within a data type, IRM or CDM, fit statistics reflect the relative diagnosticity of the data, and for fixed data, they can indicate which model is more appropriate), these statistics are not perfect indicators of how well item parameters can be estimated or attributes classified. This study shows that some parameters of items with medium and high diagnosticity can be more poorly estimated than those of less diagnostic items. In addition, these measures do not capture the qualitative difference between the data types in that CDM data, which have worse fit than their IRM counterpart, can produce lower misclassification rates. Finally, the fit statistics are global measures that cannot be expected to account for the intricate misclassification pattern across the ability or item classes.

Although the current study provides some useful insights, we acknowledge that it is limited in some ways. The limitations are related to the selective choice of measurement models, CDM and linear attribute structure, the definition of diagnosticity, and the distribution of the latent trait. As we noted earlier, our choices were pragmatically motivated to address the specific issues at hand. Consequently, the conclusions derived from this study may not necessarily apply to situations that involve different configurations. Nevertheless, the conditions created in this study represent meaningful types of diagnostic patterns a test may exhibit, and the results show the level of accuracy that can be expected under those conditions. In addition, the choice of a hierarchical linear attribute structure, although simple in terms of its complexity, has important practical and theoretical implications concerning the transition from continuous to discrete assessment frameworks. We would like to note, though, that diagnostic information based on a unidimensional latent trait or a hierarchical linear structure may not be fully diagnostic. Even if the categories defined to classify the students are interpretative, informative, and possibly prescriptive, the unidimensional constraint on these categories does not allow further differentiation of students within a category with respect to their specific strengths and weaknesses. In other words, a test designed to measure a general unidimensional latent trait is inherently less diagnostic than a test targeting the interaction of multiple finer-grained attributes.

We view this work as a first step in systematically examining the implications of retrofitting IRT data with CDMs. We note that although our current study indicates that retrofitting CDM to IRM data may not always be a prudent approach when a single latent trait is involved, these results do not preclude retrofitting of CDMs involving less structured attributes to IRM data that are not perfectly unidimensional. Data that depart from unidimensionality, if slightly, contain additional information that unidimensional IRMs cannot capture. This extra information, which would otherwise be considered noise, can have diagnostic value given the appropriate theoretical framework and CDMs.

The anticipated increase in the use of retrofitting should motivate researchers to investigate further issues surrounding this practice. It would be beneficial to study how results are affected by the use of various types of psychometric models, dimensionality of the latent trait, and structure of the attributes. The biggest challenge in this area lies in making the parameters and data from IRM and CDM framework comparable. Although this article explored methods for translating parameters between the frameworks, more research is needed in this area as well.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Addison Wesley Longman.

Birenbaum, M., Tatsuoka, C., & Xin, T. (2005). Large-scale diagnostic assessment: Comparison of eighth graders' mathematics performance in the United States, Singapore and Israel. *Assessment in Education Principles Policy and Practice*, *12*, 167–181.

Commons, M. L., & Richards, F. A. (1984). A general model of stage theory. In M. L. Commons, F. A. Richards, & C. Armon. (Eds.), *Beyond formal operations* (pp. 120–140). New York: Praeger Publishers.

de la Torre, J. (2006, June). *Skills profile comparisons at the state level: An application and extension of cognitive diagnosis modeling in NAEP*. Presentation at the International Meeting of the Psychometric Society, Montreal, Canada.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

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

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*, 285–291.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. New York: Springer-Verlag.

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

Erikson, E. H. (1950). *Childhood and society*. New York: Norton.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394–411.

Haertel, E. H. (1990) Continuous and discrete latent structure models for item response data. *Psychometrika*, *55*, 477–494.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Junker, B. W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Paper prepared for the Committee on the Foundations of Assessment, National Research Council, November 30, 1999.

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.

Kuhn, D. (2001). Why development does (and does not) occur: Evidence from the domain of inductive reasoning. In J. L. McClelland & R. Siegler. (Eds.), *Mechanisms of cognitive development: Behavioral and neural perspectives* (pp. 221–249). Hillsdale, NJ: Erlbaum.

Lee, J., Grigg, W., & Dion, G. (2007). *The nation's report card: Mathematics 2007* (NCES 2007–494). Washington, DC: National Center for Education Statistics. Retrieved April 30, 2009 from http://nces.ed.gov/nationsreportcard/itemmaps/?subj=Mathematics&year=2007&grade=4

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, *41*, 205–236.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99–120.

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

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practices. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 307–353). Washington, DC: American Educational Research Association.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Roussos, L., Templin, J., & Henson, R. (2007). Skills Diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*, 293–311.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement*, *44*, 313–324.

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. (1995). Architecture of knowledge structures and cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K., Corter, J., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal*, *41*, 901–906.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [Computer software]. Chicago: Scientific Software International.

Authors

JIMMY DE LA TORRE is Associate Professor of Educational Psychology at Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901. His areas of specialization include item response theory, cognitive diagnosis, and the use of diagnostic assessments to support classroom instruction and learning.

TZUR M. KARELITZ is Senior Research Associate, Center for Science Education, Education Development Center, 55 Chapel Street, Newton, MA, 02458; tkarelitz@edc.org. His primary research interests include assessment design and analysis.