Asia Pacific Educ. Rev. (2012) 13:333–345
DOI 10.1007/s12564-011-9196-3

Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation

Young-Sun Lee · Jimmy de la Torre · Yoon Soo Park

Received: 25 April 2011 / Revised: 20 September 2011 / Accepted: 20 October 2011 / Published online: 4 November 2011

Abstract Cognitive diagnosis models (CDMs) continue to generate interest among researchers and practitioners because they can provide diagnostic information relevant to classroom instruction and student learning. However, its modeling component has outpaced its complementary component—test construction. Thus, most applications of cognitive diagnosis modeling involve retrofitting of CDMs to assessments constructed using classical test theory (CTT) or item response theory (IRT). This study explores the relationship between item statistics used in the CTT, IRT, and CDM frameworks using such an assessment, specifically a large-scale mathematics assessment. Furthermore, by highlighting differences between tests with varying levels of diagnosticity using a measure of item discrimination from a CDM approach, this study empirically uncovers some important CTT and IRT item characteristics. These results can be used to formulate practical guidelines in using IRT- or CTT-constructed assessments for cognitive diagnosis purposes.

Keywords Cognitive diagnosis · DINA model · Item response theory · Classical test theory · Large-scale assessment

Y.-S. Lee (✉) · Y. S. Park
Teachers College, Columbia University, 525 W. 120th Street, New York, NY 10027, USA
e-mail: yslee@tc.columbia.edu

Y. S. Park
e-mail: yoon.park@gmail.com

J. de la Torre
Rutgers University, 10 Seminary Place, New Brunswick, NJ 08901, USA
e-mail: j.delatorre@rutgers.edu

Introduction

Most commonly employed unidimensional item response theory (IRT) models provide single overall scores. Although these scores can indicate students' relative position on a proficiency continuum, these models do not aim at providing more specific diagnostic information relevant to classroom instruction and student learning. In contrast, cognitive diagnosis models (CMDs) are developed specifically to identify the presence or absence of multiple fine-grained skills. A more generic term for skills, cognitive processes, knowledge states, and knowledge representations is *attributes*. Instead of a single score, a binary attribute vector is posited for each examinee to indicate which attributes the examinee has and has not mastered. Because attributes are defined at a fine-grained level, they can provide information that is more useful in practical instructional settings.

For example, consider the fraction subtraction task $2\frac{4}{12} - \frac{7}{12}$. An IRT model might describe the performance on the task as a function of fraction subtraction proficiency, and students with higher proficiencies are expected to have better chances of answering the item correctly. In contrast, a CDM might describe the performance as a function of the attributes given in Table 1.

These attributes are based on those identified by Mislevy (1995) and Tatsuoka (1990) using cognitive theory and analysis of the way a student population of interest solves this type of problem. A successful performance on the task requires a series of successful implementations of the relevant attributes. The model might also describe the implications of missing one or more of the required attributes. Thus, by incorporating cognitive structures in the psychometric model, analysis using CDMs provides information that is richer, more prescriptive, and more relevant to instruction and learning.

Table 1 Attributes for fraction subtraction

(1) Borrow one from whole number to fraction
(2) Basic fraction subtraction
(3) Reduce
(4) Separate whole number from fraction
(5) Convert whole number to fraction

For assessments to provide optimal diagnostic information, they need not only be analyzed using CDMs, and they also need to be constructed using a cognitive diagnosis framework; that is, the assessment has to be designed to provide actionable feedback as intended by CDMs. Although cognitive diagnosis continues to generate interest among researchers and practitioners, its modeling component has outpaced its complementary component—test construction. At present, very few assessments, particularly those designed for large-scale applications, are built from a cognitive diagnosis framework. Consequently, cognitive diagnosis modeling is typically applied to assessments that were constructed using a unidimensional IRT framework. The approach is referred to as *retrofitting* because CDMs are fitted to the data after the fact. Because of the original intended use of these assessments, they cannot be expected to also provide sufficient diagnostic information, at least not in their entirety. The glaring disparity between the models used in designing and analyzing an assessment casts some doubts on the diagnostic value of cognitively based analysis of IRT-based data. Unfortunately, because availability of large-scale assessments more appropriate for CDMs (i.e., assessments designed to be cognitively diagnostic) is not in the horizon, we can expect the practice of retrofitting to continue in the foreseeable future. This practice begs answers to these important questions: Is retrofitting warranted? Do IRT-constructed assessments provide diagnostic information? Should all the items be expected to be diagnostic? If not, what characteristics differentiate diagnostic items from those that are not?

The purpose of this study is to seek answers to these questions by empirically examining the relationship between CTT, IRT, and CDM statistics using a large-scale mathematics assessment and exploring whether the use of CDM item indices by retrofitting an assessment developed under an IRT framework can be used to select diagnostic items to yield an improvement in its diagnosticity. This paper is organized to present a background in theory and application of measurement models used for cognitive diagnosis. This is followed by the analysis, results, and discussion of the findings.

To accomplish the above-mentioned goals, the 2003 administration of the Florida Comprehensive Assessment Test (FCAT) was used to fit a CDM using a specification of 13 attributes. Using CDM-based item measures, the 50-item FCAT was reanalyzed by selecting 30 and 40 items, and comparing the resulting parameters to the original set using CTT and IRT indices reflecting item and person parameters. In addition, characteristics of the diagnostic and non-diagnostic items (i.e., those included and not included in the subsets) were compared in terms of their CTT difficulty and discrimination parameters and three-parameter logistic (3PL) model parameters with CDM parameters. Given the lack of assessments developed using a CDM framework, this study intends to demonstrate an empirical application of retrofitting an assessment developed from an IRT framework by formulating practical guidelines and examining whether such practice can yield diagnostic value.

The DINA model

At present, there are a few dozens of CDMS that exist in the literature (Fu and Li 2007; Rupp and Templin 2008). As a whole, these models cover a variety of the situations (i.e., types of construct, response, and dimensionality) that would be of interest to researchers in measurement, cognitive, and learning sciences. One of the models described in their survey that will be the focus of this study is the *deterministic, inputs, noisy, "and" gate* (DINA; Junker and Sijtsma 2001; de la Torre 2009) model. Compared to other CDMs, the DINA model is parsimonious and easily interpretable and requires only two parameters for each item, regardless of the number of attributes being considered. This model is appropriate in applications where the conjunction of several equally important attributes is required. Applications and discussions of the DINA model (although the model in some cases may be formulated differently) can be found in de la Torre and Douglas (2004), Doignon and Falmagne (1999), Haertel (1989), Junker and Sijtsma (2001), Macready and Dayton (1977), and Tatsuoka (2002). Due to the prevalence of CDMs, researchers are also now trying to unify similar and related CDMs. These efforts are demonstrated in the generalized DINA model (G-DINA; de la Torre 2011), the general diagnostic model (GDM; von Davier 2005), and the loglinear cognitive diagnostic model (LCDM; Burke and Henson 2008; Henson et al. 2009).

Model specification

Denote the vector of dichotomous item responses of student i to J items by Y_i, and assume that these responses are functions of the K attributes required for the test, denoted by $\alpha_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}\}$. The kth element of the attribute vector, α_{ik}, is a binary indicator representing whether the

student possesses ($\alpha_{ik} = 1$) or does not possess ($\alpha_{ik} = 0$) the kth attribute. The conditional distribution of the item response given an attribute vector is specified by the DINA model (see below for details).

Like many CDMs, implementation of the DINA model requires construction of a Q-matrix (Tatsuoka 1983). A Q-matrix is a $J \times K$ matrix of zeros and ones, where the jkth element of the matrix, q_{jk}, indicates whether the kth attribute is necessary for correctly answering the jth item. Maximizing the probability of a correct response to item j necessitates that the students possess all the required attributes specified in the jth row of the Q-matrix.

Given the ith student's attribute vector α_i and the jth row of the Q-matrix (i.e., the row corresponding to the jth item), a latent response η_{ij} is generated deterministically through the equation $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. The latent response η_{ij} assumes the value one or zero, where $\eta_{ij} = 1$ indicates that student i has all the attributes required for item j, and $\eta_{ij} = 0$ indicates that student i lacks at least one of the attributes required for item j. Equivalently, the latent response η_{ij} can also be viewed as the examinee's latent group membership with respect to his or her mastery or lack of the required attributes for item j. However, due to the stochastic nature of the processes involved in any assessments, a latent response of 1 does not guarantee a correct observed response, nor does a latent response of zero guarantee an incorrect observed response. The stochastic nature of the model allows for the possibility that students in group 1 (i.e., $\eta_{ij} = 1$) may slip and answer the item incorrectly, while students in group 0 (i.e., $\eta_{ij} = 0$) may guess and answer the item correctly. It should be noted that guessing in the context of this model is not limited to correct response by chance, but may include reliance on strategies not specified in the Q-matrix (de la Torre and Douglas 2008; Maris 1999).

The probability that a student in group 1 (i.e., a student who possesses all the required attributes) slips and incorrectly answers an item j is denoted by the slip parameter s_j; the probability that a student in group 0 (i.e., a student who lacks at least one of the required attributes) guesses and correctly answers an item j is denoted by the guessing parameter g_j. Two parameters in the DINA model, the *slip* and *guessing* parameters, can be expressed as $s_j = P(X_{ij} = 0|\eta_{ij} = 1)$ and $g_j = P(X_{ij} = 0|\eta_{ij} = 1)$ for item, respectively (de la Torre 2009). Given s_j and g_j, the item response function that relates the conditional distribution of the response to item j and the attribute vector α_i can be defined as.

$$P(Y_{ij} = 1|\alpha_i) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \quad (1)$$

Estimation

Several methods of estimating the DINA model parameters exist. One approach is to treat the DINA model like any traditional latent variable model (i.e., IRT) and to estimate its parameters using the marginalized maximum likelihood estimation (MMLE) method (Bock and Aitkin 1981). A detailed description of this method as applied to the DINA model can be found in de la Torre (2009). MMLE estimation requires integration over the distribution of the latent variable. In cognitive diagnosis modeling, this refers to the distribution of all the possible patterns of the attributes. When K is large, say, at least 20, the number of possible attribute patterns may render MMLE virtually impossible to implement. To overcome this problem, de la Torre and Douglas (2004) introduced an approach that allows the distribution of the attribute patterns to be simplified. Specifically, the number of parameters to be estimated can be greatly reduced by assuming that the attributes are conditionally independent given a higher-order latent trait. The model parameters under this framework have been estimated using Markov chain Monte Carlo methods.

Discrimination index

de la Torre (2007) proposed a DINA-based item discrimination index defined as $\delta_j = 1 - s_j - g_j$. To illustrate the interpretation of this index, two hypothetical items with different discrimination indices are given in Fig. 1. Graphically, the differences between the heights of the probabilities of a correct response between $\eta_{ij} = 0$ and $\eta_{ij} = 1$ represent the discrimination indices of the items. Compared to the item in Panel A, the item in Panel B provides greater differentiations between the two groups of examinees.

Another way of interpreting δ_j is in terms of the posterior probability of an examinee's group membership with respect to an item conditional on the observed response to the item. That is, the item discrimination in the DINA context can be interpreted as an indicator of the strength of the relationship between an examinee's group membership and observed response such that inferring about an examinee's latent group membership based on the observed response is less unequivocal when the discrimination index is higher.

For a correct observed response, it is of interest to compare the following posterior probabilities: $P(\eta_{ij} = 0|Y_{ij} = 1)$ and $P(\eta_{ij} = 1|Y_{ij} = 1)$. The posterior probability of being in group 1 given a correct observed response can be written as

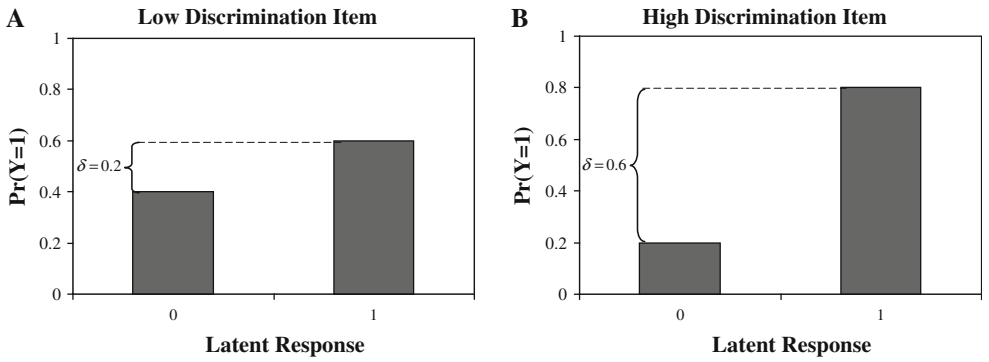


Fig. 1 Two hypothetical items with different discrimination indices

$$P(\eta_{ij}=1|Y_{ij}=1)=\frac{P(Y_{ij}=1|\eta_{ij}=1)P(\eta_{ij}=1)}{P(Y_{ij}=1|\eta_{ij}=0)P(\eta_{ij}=0)+P(Y_{ij}=1|\eta_{ij}=1)P(\eta_{ij}=1)}. \quad (2)$$

For illustration purposes, assume that the prior probabilities of group memberships are equal—$P(\eta_{ij}=0)=P(\eta_{ij}=1)=0.5$. The posterior probability in (2) simplifies to

$$\begin{aligned}P(\eta_{ij}=1|Y_{ij}=1)&=\frac{P(Y_{ij}=1|\eta_{ij}=1)}{P(Y_{ij}=1|\eta_{ij}=0)+P(Y_{ij}=1|\eta_{ij}=1)}\\&=\frac{1-s_j}{1-s_j+g_j}.\end{aligned} \quad (3)$$

Similarly, the posterior probability of being in group 0 given a correct observed response can be written as

$$\begin{aligned}P(\eta_{ij}=0|Y_{ij}=1)&=\frac{P(Y_{ij}=1|\eta_{ij}=0)}{P(Y_{ij}=1|\eta_{ij}=0)+P(Y_{ij}=1|\eta_{ij}=1)}\\&=\frac{g_j}{1-s_j+g_j}.\end{aligned} \quad (4)$$

For the inference about an examinee's latent group membership to be less unequivocal, the difference between the two posterior probabilities needs to be maximized. That is, we take

$$\begin{aligned}P(\eta_{ij}=1|Y_{ij}=1)-P(\eta_{ij}=0|Y_{ij}=1)&=\frac{1-s_j-g_j}{1-s_j+g_j}\\&=\frac{\delta_j}{1-s_j+g_j}.\end{aligned} \quad (5)$$

It can be noted that when $P(\eta_{ij}=0)=P(\eta_{ij}=1)$, $(1-s_j+g_j)/2$ represents the expected proportion of examinees answering item j correctly. Thus, for a fixed level of item difficulty (i.e., $1-s_j+g_j=c$), the difference between the posterior probabilities is maximized by maximizing δ_j.

Given in Table 2 are different guessing and slip parameters for an item with 0.4 probability of being answered correctly (i.e., $1-s_j+g_j=0.8$). The discrimination index and posterior probabilities of being in groups 0 and 1 given a correct response are also tabulated. In addition to the assumption that the prior group memberships are equal, these entries also assume that the probability that an examinee from group 0 will give a correct response cannot be greater than that of group 1. This table shows that the item has the lowest discrimination index when $g_j=0.4$ and $s_j=0.6$, which implies that both groups have a 0.4 probability of answering the item correctly and $\delta_j=0$.

Consequently, a correct response provides no indication as to the examinee's group membership. In comparison, when $g_j=0.1$, $s_j=0.3$, $\delta_j=0.6$, and there is a 0.875 probability, an examinee who gave a correct response is from group 1; this probability and δ_j achieve their highest values, 1.0 and 0.8, respectively, when $g_j=0.0$ and $s_j=0.2$. It should be noted that, in general, there is no one-to-one correspondence between δ_j and the conditional probabilities. That is, items with the same discrimination index but different levels of difficulty result in different posterior probabilities. Although posterior probabilities may be more informative in that they take into account both the item discrimination and difficulty, their computation is not straightforward (i.e., they require the availability of prior information about the latent group memberships which is not known and may vary from one population to another). In contrast, estimates of δ_j can be obtained directly from the estimates of guessing and slip parameters of the DINA model.

Table 2 Posterior probabilities of group membership and δ_j for different s_j and g_j : $P(\eta_{ij} = 0) = P(\eta_{ij} = 1)$ and $1 - s_j + g_j = 0.8$

g_j	s_j	δ_j	Posterior probabilities	
			$\eta_{ij} = 0$	$\eta_{ij} = 1$
0.400	0.600	0.000	0.500	0.500
0.200	0.400	0.400	0.250	0.750
0.100	0.300	0.600	0.125	0.875
0.050	0.250	0.700	0.063	0.938
0.025	0.225	0.750	0.031	0.969
0.000	0.200	0.800	0.000	1.000

Method

Data

Data were taken from the 2003 administration of the Florida Comprehensive Assessment Test (FCAT; Florida Department of Education 2003a) 9th grade Mathematics test. The FCAT measures achievement in reading and mathematics for Florida students in Grades 3 through 10, in science at Grades 5, 8, and 11, and in writing at Grades 4, 8, and 10 by assessing student progress on benchmarks identified in the *Sunshine State Standards*. The FCAT was developed using an IRT framework; that is, items were calibrated, scaled, and equated using the 3PL IRT model (Turhan 2006; Florida Department of Education 2003b). The 9th grade mathematics test consists of 50 multiple choice items and assesses five contents: Number Sense, Concepts, and Operations (NS, 18%); Measurement (M, 18%); Geometry and Spatial Sense (GS, 24%); Algebraic Thinking (AT, 22%); and Data Analysis and Probability (DP, 18%). Items were dichotomously scored, and any items that were omitted or not reached were scored as incorrect. A sample of 1,500 examinees was randomly drawn from the sample of about 135,800 who took the test.

Analysis

The set of attributes used in this study was adopted from the National Council of Teachers of Mathematics (NCTM) *Principles and Standards for School Mathematics* (NCTM 2000). Attributes were classified into 5 contents, and the list of the attributes is given in Table 3, which also includes the number of times the attributes were specified. A Q-matrix was specifically constructed for the data. Two secondary-school mathematics teachers who are graduate students in mathematics education and have experience teaching of 9th grade mathematics, and one researcher who

Table 3 Attributes adopted from NCTM principles and standards to analyze performance on mathematics items from the FCAT (2003) for 9th grade

Strand	Attributes	Number of times specified		
		50-item	40-item	30-item
Number sense, concepts, and operations (NS)	Understand numbers, ways of representing numbers, relationships among numbers, and number systems	12	8	7
	Understand meanings of operations and how they relate to one another	15	12	5
	Compute fluently and make reasonable estimates	32	28	20
Measurement (M)	Understand measurable attributes of objects and the units, systems, and processes of measurement/apply appropriate techniques, tools, and formulas to determine measurements	13	12	11
Geometry and Spatial Sense (GS)	Analyze characteristics and properties of two- and three-dimensional geometric shapes and develop mathematical arguments about geometric relationships/specify locations and describe spatial relationships using coordinate geometry and other representational systems	11	8	6
	Apply transformations and use symmetry to analyze mathematical situations	15	13	12
	Use visualization, spatial reasoning, and geometric modeling to solve problems	19	13	10
Algebraic thinking (AT)	Understand patterns, relations, and functions/analyze changes in various context	2	2	1
	Represent and analyze mathematical situations and structures using algebraic symbols	13	10	6
	Use mathematical models to represent and understand quantitative relationships	23	19	12
Data analysis and probability (DP)	Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them	3	2	1
	Select and use appropriate statistical methods to analyze data/develop and evaluate inferences and predictions that are based on data	6	4	3
	Understand and apply basic concepts of probability	5	4	3

is a domain expert identified the attributes requirements for each of the items to be solved correctly. Each coder solved the items independently and, if necessary, discussed the discrepancies until consensus was reached. Also if items can be solved using different strategies, the most dominantly employed strategy was used to define the Q-matrix. This process ensured the validity and usefulness of the attributes and the Q-matrix. We constructed or adopted four items (see Fig. 2) from the released FCAT to illustrate typical items in the test. Also given in the figure are the attribute specifications for the items. It should be noted that although this Q-matrix, as well as attributes, represents one of several possibilities, the validation process involving subject matter experts affords it theoretical grounding and practical utility. However, refinement based on further research and analysis may be needed to empirically correct possible misspecifications (de la Torre 2008).

Given the Q-matrix, the data were analyzed using an MMLE of the DINA model parameters. The computer program Ox (Doornik 2002) was used to implement the algorithm. The console version of Ox is available free of charge for academic research and teaching purposes, and the MMLE code for the DINA model used in this study can be made available upon request. Based on $\hat{g}$ and $\hat{s}$ of the entire assessment (i.e., 50-item test), the DINA-based discrimination index (i.e., $\hat{\delta}_j = 1 - \hat{s}_j - \hat{g}_j$) was computed. Two subsets of the assessment, representing the best possible 30- and 40-item tests in that they consisted of items with the highest $\hat{\delta}$, were constructed. The three-parameter logistic model analysis of the data using BILOG-MG (Zimowski et al. 1996) was also carried out. The item information functions (IIFs) and item characteristic curves (ICCs) were computed using the item parameter estimates, $\hat{a}_j, \hat{b}_j$, and $\hat{c}_j$. A CTT analysis of the data was conducted. Specifically, the items were described using CTT indices, namely proportion correct or item easiness, p_j+, and item discrimination d_j, where $p_j+ = \sum_{i=1}^{1,500} Y_{ij} \Big/ 1,500$ and $d_j = \mathrm{Cor}\left(Y_{ij}, \sum_{j=1}^{50} Y_{ij} - Y_{ij}\right)$.

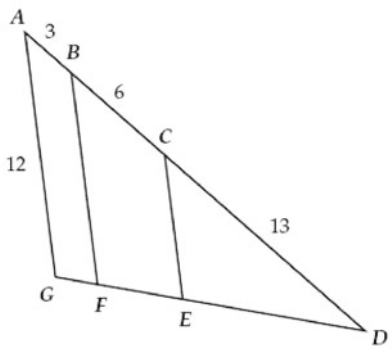
Fig. 2 Four sample test items and their corresponding Q-vectors

Example 1. Houses of three friends of Jen – Ace, Bryan, and Cindy – are north, east and west of Jen's house. Ace's house is 3×10^3 units north from Jen's and Bryan's house is 4×10^3 units east from Jen's. If Cindy's house is west to the Jen's house with the same distance of Ace's and Bryan's houses, then how far is Cindy's house from Bryan's house?

a. 5×10^3 unit b. 7×10^3 unit c. 8×10^3 unit d. 9×10^3 unit

NS1	NS2	NS3	M1	GS1	GS2	GS3	AT1	AT2	AT3	DP1	DP2	DP3
1	1	1	0	0	1	1	0	0	0	0	0	0

Example 2. In triangle *ADG* below, the length of side *DG* is 18 units. Line segments *AG*, *BF*, and *CE* are all parallel.



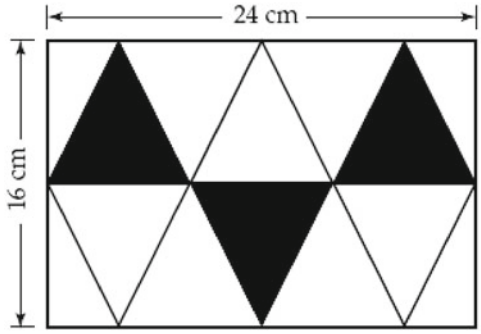
What is the approximate length of line segment *EG* ?

a. 4.9 units b. 7.4 units c. 11.0 units d. 12.5 units

NS1	NS2	NS3	M1	GS1	GS2	GS3	AT1	AT2	AT3	DP1	DP2	DP3
0	0	1	0	0	0	1	1	1	0	0	0	0

Fig. 2 continued

Example 3. An artist is designing a pattern of triangular tiles to cover a wall. Each section of the pattern is identical to the section shown below.



If the wall is 576 cm long and 384 cm high, how many black tiles will the artist need to use on the wall?

a. 72 b. 576 c. 1,152 d. 1,728

NS1	NS2	NS3	M1	GS1	GS2	GS3	AT1	AT2	AT3	DP1	DP2	DP3
0	1	1	1	1	1	0	0	0	0	0	0	0

Example 4. At a fast food restaurant, there are 4 different types of sandwiches are available: Italian, BLT, tuna melt, and cheese stake. A customer wants to order one 6-inch and one 9-inch of different kind. How many combinations of two sizes of sandwiches are possible?

a. 4 b. 6 c. 8 d. 16

NS1	NS2	NS3	M1	GS1	GS2	GS3	AT1	AT2	AT3	DP1	DP2	DP3
0	0	1	0	0	0	0	0	0	0	1	1	1

Results

The distribution of scores in terms of percent correct is illustrated in Fig. 3. This shows a balance in the range of possible scores in the FCAT assessment (Mean = 47.61; SD = 20.42).

From Table 3, the percent of times the attributes were specified by item strand in the 50-item test were as follows: NS (attributes 1–3) 35%, M (attribute 4), 8%, GS (attributes 5–7) 27%, AT (attributes 8–10) 22%, and DP (attributes 11–13) 8%. The distributions of the percent of times the attributes were specified for the 30- and 40-item tests were only slightly different. In terms of content strands, Fig. 4 shows using δ to reduce the number of item did not result in tests that have dramatically different compositions. That is, the relative proportions of items from the different content strands were comparable across the three test lengths.

Table 4 presents the estimates of $p+$, d, a, b, c, g, $1 - s$, and δ for all the items based on the analysis of the 50-item test. Some numbers were followed by the symbols * and ** to indicate which items comprised the different tests. In comparing the first and second halves of the full-length test, there was no indication that items deleted to create the 30- and 40-item tests were systematically related to their positions in the test. That is, 6 and 4 items of the first 10 deleted items were from the first and second halves of the test, respectively; these numbers were 4 and 6 for the second 10 items deleted.

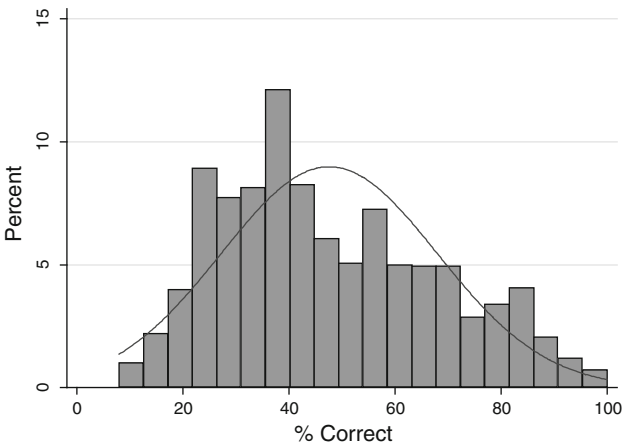


Fig. 3 Distribution of percentage of correct scores

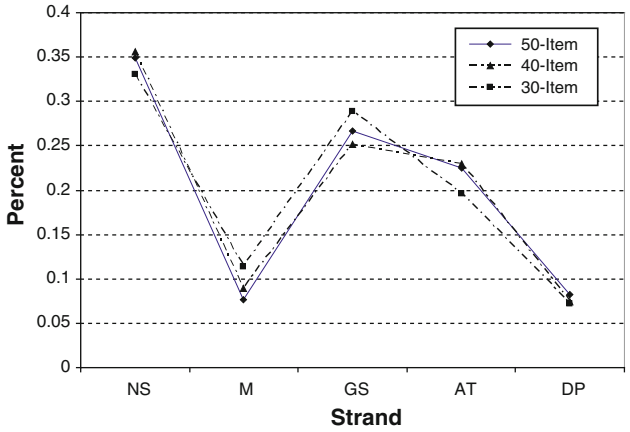


Fig. 4 Percentage of item strand by test length

Table 4 also reveals that of the 10 items with the lowest δ, four items (1, 2, 3, and 26) had some of the highest g and $p+$, and lowest b; among these 10 items, items 2, 3, 9, and 26 had some of the highest c. It also shows that seven items with the smallest d (1, 9, 16, 26, 39, 41, and 47) were also among the 10 items with the smallest δ. These results strongly suggest that the DINA indices are related to the IRT and CTT indices.

To further explore the relationships between the DINA, IRT, and CTT indices, the correlations between these indices were computed (see Table 5). The correlation table shows high absolute correlations of $p+$ and b with g and $1 - s$, and d with δ. In addition, a and c were moderately correlated with g, and c with δ. These results indicate that items with low discrimination d, and to some extent, high guessing parameter c, can be expected to have low discrimination index δ.

A graphical illustration of the relationship between CTT and DINA indices given in Fig. 5 shows that the first 10 items deleted were either among the items with lowest d or highest $p+$. In contrast, items included in the 30-item test were of at least moderate discrimination and were not extremely easy. The plot also shows that the second set of 10 deleted items had a wider range of $p+$ and lower overall discrimination compared to the remaining 30 items. Finally, 13 of the 20 deleted items were close to the mean difficulty level or below it.

Table 4 CTT, 3PL, and DINA item parameter estimates

Item	Strand	CTT		3PL			DINA		
		$p+$	d	a	b	c	g	$1 - s$	δ
1**	DP	0.94	0.24	1.65	−2.22	0.18	0.87	1.00	0.13
2**	DP	0.84	0.36	1.55	−1.34	0.13	0.72	0.99	0.27
3**	AT	0.76	0.31	1.01	−1.18	0.18	0.65	0.92	0.26
4	AT	0.61	0.36	1.10	−0.15	0.17	0.44	0.82	0.38
5*	NS	0.68	0.39	1.17	−0.60	0.15	0.56	0.89	0.33
6*	NS	0.41	0.29	1.64	1.24	0.10	0.30	0.62	0.32
7*	AT	0.83	0.38	1.70	−1.25	0.12	0.69	0.99	0.30
8	AT	0.55	0.39	1.11	−0.06	0.11	0.41	0.78	0.37
9**	GS	0.62	0.23	0.64	−0.24	0.34	0.55	0.75	0.19
10	NS	0.47	0.54	1.92	0.28	0.06	0.21	0.84	0.62
11*	GS	0.35	0.28	1.74	1.40	0.09	0.26	0.55	0.29
12	DP	0.68	0.31	0.97	−0.46	0.25	0.51	0.88	0.37
13	DP	0.26	0.60	2.23	0.86	0.04	0.05	0.64	0.59
14	GS	0.47	0.55	2.25	0.31	0.06	0.24	0.84	0.60
15	GS	0.55	0.50	1.66	−0.03	0.07	0.25	0.96	0.70
16**	M	0.33	0.19	3.49	1.52	0.07	0.27	0.45	0.18
17	M	0.56	0.51	2.16	0.14	0.07	0.36	0.89	0.54
18**	NS	0.40	0.31	1.66	1.21	0.09	0.30	0.59	0.29
19	M	0.18	0.58	2.70	1.14	0.04	0.02	0.49	0.47
20	GS	0.34	0.56	2.26	0.67	0.04	0.13	0.72	0.58
21	GS	0.36	0.64	2.66	0.50	0.03	0.11	0.79	0.68
22	M	0.43	0.46	2.73	0.71	0.05	0.23	0.74	0.51
23	AT	0.49	0.40	1.57	0.51	0.10	0.23	0.74	0.51
24	NS	0.53	0.46	2.52	0.44	0.06	0.33	0.84	0.51
25	AT	0.58	0.41	1.73	0.27	0.10	0.38	0.85	0.47

Table 4 continued

Item	Strand	CTT		IRT			DINA		
		$p+$	d	a	b	c	g	$1-s$	δ
26**	GS	0.88	0.28	1.20	−1.71	0.25	0.74	0.99	0.24
27	DP	0.61	0.48	1.70	−0.17	0.09	0.34	0.96	0.62
28*	AT	0.45	0.30	2.50	1.10	0.07	0.31	0.64	0.33
29*	DP	0.39	0.31	0.97	1.05	0.14	0.26	0.60	0.34
30	NS	0.44	0.53	1.98	0.43	0.06	0.21	0.81	0.60
31	M	0.36	0.55	2.16	0.62	0.05	0.18	0.74	0.57
32	NS	0.52	0.30	1.42	0.84	0.13	0.38	0.72	0.35
33	DP	0.37	0.33	1.38	1.15	0.10	0.18	0.56	0.38
34	AT	0.30	0.38	2.00	1.24	0.07	0.18	0.56	0.39
35*	M	0.20	0.31	2.05	1.63	0.08	0.11	0.42	0.32
36	GS	0.25	0.61	2.61	0.87	0.04	0.05	0.64	0.59
37	GS	0.32	0.52	1.76	0.78	0.05	0.08	0.60	0.52
38	DP	0.50	0.40	1.16	0.21	0.10	0.20	0.75	0.56
39**	NS	0.38	0.25	1.36	1.54	0.13	0.29	0.51	0.22
40*	GS	0.49	0.30	1.37	0.94	0.12	0.37	0.70	0.32
41**	M	0.23	0.24	1.71	1.83	0.12	0.16	0.38	0.22
42	AT	0.43	0.51	1.80	0.47	0.06	0.24	0.77	0.53
43	GS	0.56	0.46	1.53	0.01	0.10	0.32	0.83	0.51
44	AT	0.62	0.42	1.26	−0.33	0.10	0.29	0.84	0.55
45*	GS	0.16	0.44	1.80	1.50	0.07	0.05	0.40	0.35
46	M	0.32	0.54	1.88	0.72	0.04	0.11	0.66	0.56
47**	AT	0.20	0.23	2.79	1.72	0.08	0.09	0.30	0.21
48	NS	0.48	0.41	1.73	0.62	0.08	0.26	0.74	0.48
49	DP	0.67	0.44	1.54	−0.46	0.11	0.42	0.94	0.52
50*	M	0.49	0.32	1.55	0.88	0.11	0.36	0.71	0.35

*Deleted from 30-item test; **deleted from 30- and 40-item tests

Table 5 Correlation between CTT, IRT, and DINA statistics

	d	a	b	c	g	$1-s$	δ
$p+$	−0.24	−0.52	−0.93	0.62	0.94	0.87	−0.20
d		0.35	−0.02	−0.63	−0.48	0.23	0.90
a			0.43	−0.71	−0.51	−0.36	0.25
b				−0.51	−0.83	−0.89	0.03
c					0.73	0.30	−0.60
g						0.69	−0.50
$1-s$							0.29

The relationships between the DINA and 3PL indices are shown in Fig. 6 and provide similar results as the CTT figure. The plots show that the first 10 items deleted from the 50-item test tended to have high positive b, or relatively low a and b, and high c. Thirteen of the 20 deleted items had higher positive b (i.e., more difficult) than the average item. Individually, the a and c of the 10 deleted items were not too different from those of the last 30 items. However, taking the two-item parameters in concert, those included in the 30-item test had a tendency to have higher a and lower c as the items became more difficult.

The summary of the CTT, 3PL, and DINA indices for the different test lengths is given in Table 6. It is noteworthy that the 30-, 40-, and 50-item tests had comparable Cronbach's alpha estimates of reliability, 0.91 versus 0.92. On the average, the 30-item test had g, $1-s$, b, c, and $p+$ that were lower, and δ, a, and d that were higher. In addition, the indices of the shorter test were more uniform (i.e., the indices had lower variability). It can be noted that the mean $p+$ indicated that the 30-item test was more difficult than the 50-item test, whereas the mean b indicated the opposite. These results can be explained by the fact that the CTT difficulty parameter is bounded (i.e., between 0 and 1), but its IRT counterpart is not. Consequently, the mean of b was subjected to a higher degree of fluctuation as extremely easy or difficult items were deleted from the test.

In comparison, the impact of extremely easy and difficult items on $p+$ was not as pronounced. For this reason, the CTT index is a more reliable measure of relative item

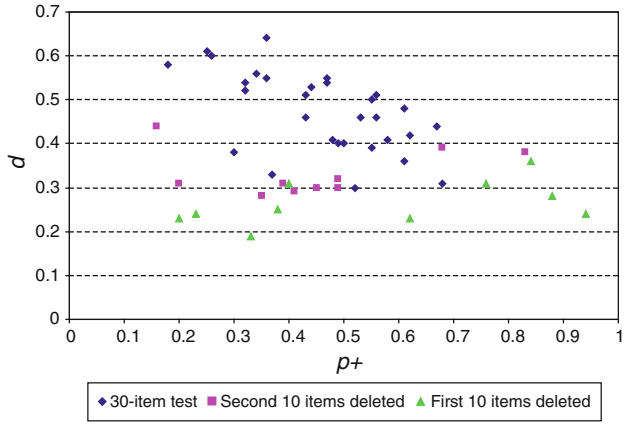


Fig. 5 CTT statistics scatter plot

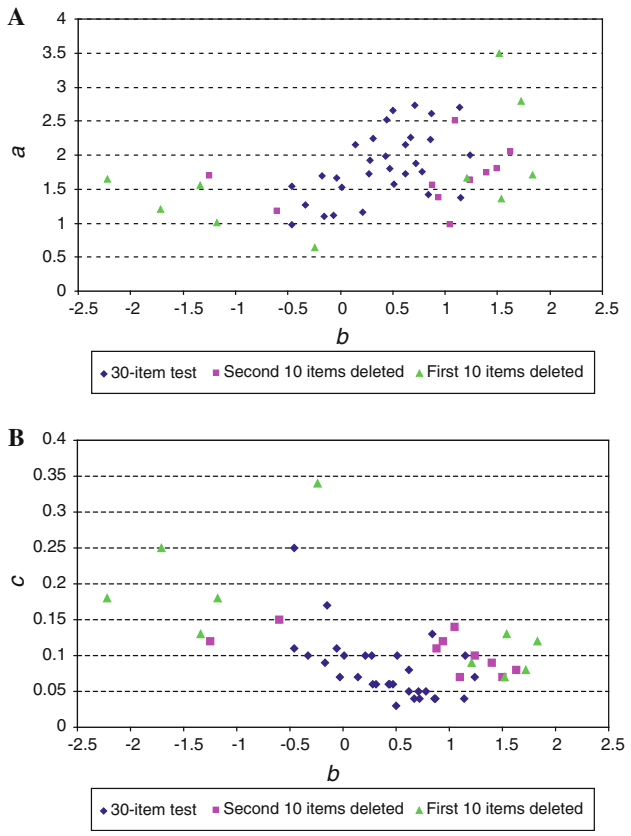


Fig. 6 Scatter plot of IRT parameters. **a** Difficulty versus discrimination. **b** Difficulty versus guessing

difficulty. In comparing the 40- and 50-item tests, the same results, albeit less pronounced, were obtained, except for the mean of *b* which was higher for the 40-item test. These findings indicate that not all the items in the original assessment were equally cognitively diagnostic. For an assessment of comparative length to be more diagnostically informative, a test with items that are slightly more

Table 6 Mean indices (SD) and reliability by test length

Statistics	50-Item test	40-Item test	30-Item test
$p+$	0.48 (0.18)	0.46 (0.15)	0.46 (0.13)
d	0.40 (0.12)	0.44 (0.10)	0.47 (0.09)
a	1.78 (0.57)	1.80 (0.49)	1.85 (0.50)
b	0.42 (0.90)	0.50 (0.63)	0.40 (0.47)
c	0.10 (0.06)	0.09 (0.04)	0.08 (0.05)
g	0.31 (0.19)	0.27 (0.15)	0.24 (0.13)
$1 - s$	0.73 (0.18)	0.74 (0.14)	0.76 (0.12)
δ	0.42 (0.15)	0.47 (0.12)	0.52 (0.09)
Reliability (Cronbach's α)	0.92	0.91	0.91

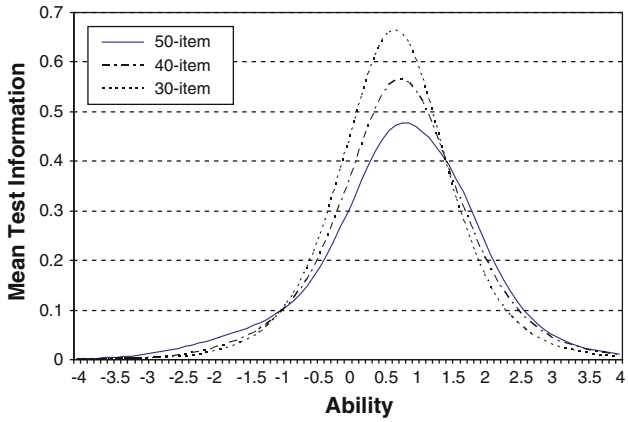


Fig. 7 Mean item information function

homogeneous, and have lower $p+$ and higher d (using CTT statistics), or moderate b with lower c and higher a (using IRT statistics) is needed.

Figures 7 and 8 give the mean IIFs and ICCs for the three test lengths. The mean IIF of the 30-item test was higher in the interval -1.0 through about 1.5, indicating that it is most informative in this region; in contrast, the longer tests were more informative in the regions beyond.

The mean ICCs show that the 30-item test was relatively harder for students with abilities less than 0.5, whereas the longer tests were more difficult for examinees with higher abilities. Therefore, in addition to the above characteristics, more diagnostically informative test appeared to be more discriminating where most of the examinees were located and more difficult for examinees with low-to-medium high abilities.

With respect to person parameter estimates, Table 7 shows correlation between DINA posterior probability of attribute mastery with CTT percent correct and IRT ability estimates (θ). Posterior probabilities were computed for the 30-, 40-, and 50-item tests, whereas percent correct and ability estimates were computed only for the 50-item test.

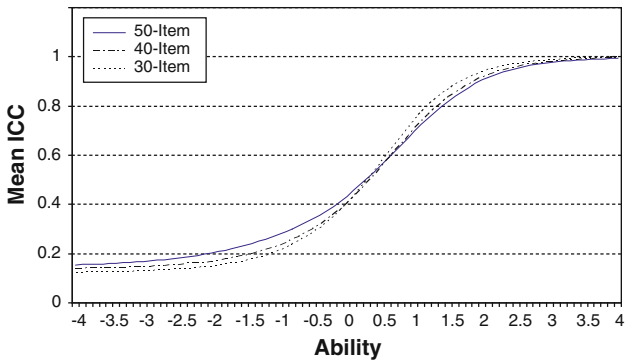


Fig. 8 Mean item characteristic curves

Across the three test lengths, the correlations were similar. Although there were small variations in the relationship across test lengths, this indicates that even with the reduction in items using the DINA item discrimination index, person parameters were comparable.

Table 8 presents the model fit indices of the IRT and DINA models. The IRT model indices are presented for the 50-item assessment, while this is contrasted with the 30-, 40-, and 50-item assessment using the DINA model. Based on the deviance (-2LL) estimates, results show that 50-item 3PL IRT model fits better than the 50-item DINA model.

Discussion and conclusion

Retrofitting assessments constructed using a unidimensional IRT or CTT framework with CDMs has become a commonplace in the literature; in fact, there has been an increase in the application of retrofitting international assessments. Studies such as Lee et al. (2009) have examined the utility of retrofitting the Trends in International Mathematics and Science Study (TIMSS) to investigate the diagnostic value of CDM analysis for students in Korea and the U.S; similar to the FCAT, TIMSS was developed under a unidimensional IRT framework. Yet, studies have shown that the CDM approach can yield rich diagnostic information as well as produce comparable model fit (Lee et al. 2011).

Table 7 Correlations of Pearson estimates: 30-, 40-, and 50-item DINA posterior probabilities with 50-item CTT percent correct and 50-item IRT ability estimates

Attribute	30-items DINA		40-items DINA		50-items DINA	
	CTT % correct	IRT theta	CTT % correct	IRT theta	CTT % correct	IRT theta
1	0.80	0.80	0.82	0.80	0.76	0.79
2	0.71	0.69	0.76	0.72	0.64	0.58
3	0.75	0.76	0.81	0.81	0.81	0.81
4	0.79	0.80	0.78	0.78	0.78	0.81
5	0.78	0.77	0.72	0.67	0.66	0.59
6	0.79	0.77	0.73	0.68	0.81	0.78
7	0.77	0.76	0.81	0.82	0.74	0.74
8	0.51	0.49	0.51	0.49	0.61	0.58
9	0.75	0.71	0.77	0.72	0.76	0.71
10	0.74	0.72	0.76	0.72	0.74	0.69
11	0.58	0.57	0.74	0.71	0.68	0.70
12	0.67	0.63	0.69	0.62	0.69	0.67
13	0.78	0.79	0.69	0.65	0.78	0.79

Table 8 IRT and CDM model fit indices

	3PL IRT 50 items	DINA		
		30 items	40 items	50 items
−2 LL	80,247	50,331	67,293	81,969

Although such practice has not become operational in real-world settings, retrofitting has an implicit underlying assumption that items can provide diagnostic information even if they are originally developed for a different purpose. As evidence from Table 8, between the 50-item DINA and the 50-item 3PL IRT models, fit indices favor the latter model. This may be an expected result, as the FCAT was developed under an IRT framework (de la Torre and Karelitz 2009). However, the results also demonstrate that using fewer items (e.g., 30 items) through the use of discrimination indices in the DINA model can provide a better model fit that may be more diagnostic in assessing examinees' skills. As such, there can be a trade-off in that the attribute specification that is required in CDM can be less parsimonious when compared to the simpler unidimensional IRT model. In addition, it also assumes that these items are of equal diagnostic value. To investigate the relationships between the CTT, IRT, and CDM frameworks, a large-scale assessment was analyzed using the different frameworks, and the resulting CTT, 3PL, and DINA indices were compared. By highlighting the differences between tests consist of items with varying level of diagnosticity, some important CTT and IRT item characteristics as they relate to cognitive diagnosis were empirically uncovered.

Results of this study show that, although the IRT- or CTT-based test can be diagnostically informative, the items in the test can provide different levels of diagnostic information. Specifically, items that are diagnostic are slightly less difficult but highly discriminating in the CTT sense, or more moderately difficult with lower guessing and higher discrimination parameters in the IRT sense. In addition, as a whole, more discriminating tests have less variable CTT and IRT statistics and thus provide more accurate measurement of examinees in the middle range of the ability continuum. These results can be a useful test

construction tool in developing cognitively diagnostic tests from a CTT or IRT perspective.

Another important practical finding of this study pertains to the CTT reliabilities of shorter tests—it shows that by using the DINA discrimination index, as much as 40% of the items can be deleted without adverse consequence to the test reliability. One can capitalize on this result and build a test of the same length (i.e., 50 items), but includes items are more informative in other respects. For example, the first 30 items may be retained to satisfy the CTT requirement of the test; however, the last 20 items can be strategically selected to maximize the diagnostic information of the test or to measure specific regions of the ability continuum more accurately.

This study has shown that IRT- or CTT-constructed assessments can have diagnostic value. In addition, it also provides some guidelines into how diagnostic information from such assessment can be improved without using an assessment developed under a fully cognitive diagnostic framework. However, it should be noted that because assessments retrofitted with CDMs serve multiple purposes, they necessarily cannot provide optimal diagnostic information. For assessment to be most useful diagnostically, test construction must follow from the analysis framework, rather than the other way around (Mislevy 1994, 1995). That is, instead of deriving the definitions of the attributes from existing items, items must be constructed based on the specific definitions of the attributes adopted for a particular domain. These definitions, in turn, must reflect the most current understanding of how students think and learn. Thus, a more complete implementation of cognitive diagnosis modeling goes beyond model fitting but also involves appropriate development of test items and incorporation of recent developments in other disciplines such as cognitive and learning sciences, an ideal articulated by the National Research Council (2001).

The results of this study are based on a particular assessment and framework of analysis. As such, this study has several limitations that one needs to consider in generalizing its results. First, the results may not hold in their entirety when applied to a different domain or group of examinees. Second, the inferences may change as the number of attributes and their definitions are changed. Third, even with the same assessment, set of attributes, and group of examinees, the findings may differ from the current findings if a different CDM is employed. These limitations point to the different directions one can take to extend the present study. It would be worthwhile to investigate whether the use of different data from a different domain using a different set of attributes and CDM would yield similar conclusions found in this study. Furthermore, the use of simulations through known parameter values can aid in understanding how the different measurement frameworks function to supplement empirical findings.

References

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

Burke, M. J., & Henson, R. (2008). *LCDM user's manual*. Greensboro: University of North Carolina at Greensboro.

de la Torre, J. (2007, April). *Evaluation of model fit in a large-scale assessment application of cognitive diagnosis*. Presentation at the annual meeting of the national council on measurement in education, Chicago, IL.

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. (2009). 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., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353.

de la Torre, J., & Douglas, J. (2005, April). *Modeling multiple strategies in cognitive diagnosis.* Presentation at the annual meeting of the national council on measurement in education, Montreal, Canada.

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

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

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

Doornik, J. A. (2002). *Object-oriented matrix programming using Ox* (Version 3.1). [Computer software]. London: Timberlake Consultants Press.

Florida Department of Education. (2003a). *Florida comprehensive assessment test.* Tallahassee, FL: Author.

Florida Department of Education. (2003b). *Florida comprehensive assessment test for reading and mathematics: Technical report for test administrations of FCAT 2003.* San Antonio, TX: Harcourt Educational Measurement.

Fu, J., & Li, Y. (2007, April). *An integrative review of cognitively diagnostic psychometric models.* Presentation at the annual meeting of the national council on measurement in education, Chicago, IL.

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

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

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

Lee, Y.-S., Choi, K.-M., & Park, Y. S. (2009, April). *A comparison between the U.S. and Korea in the TIMSS 8th grade mathematics assessment: An application of cognitive diagnostic modeling*.

Presentation at the annual meeting of the american education research association, San Diego, CA.

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). An analysis of attribute mastery via cognitive diagnostic modeling: A comparison of Massachusetts, Minnesota, and the U.S. national average via TIMSS 2007 4th grade mathematics. *International Journal of Testing, 11*(2), 144–177.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 33*, 379–416.

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

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439–483.

Mislevy, R. (1995). Probability-based inference in cognitive diagnosis. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 43–71). Hillsdale: Erlbaum.

National Council of Teachers of Mathematics (NCTM). (2000). *Principles and standards for school mathematics*. Reston: NCTM.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington: National Academies Press.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement, 6*, 219–262.

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

Tatsuoka, K. K. (1990). Toward an integration of item–response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale: Erlbaum.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C (Applied Statistics), 51*, 337–350.

Turhan, A. (2006). *Multilevel 2PL item response model vertical equating with the presence of differential item functioning*. Unpublished doctoral dissertation, The Florida State University, Tallahassee, FL.

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

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